

Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis

Joao Gabriel Rodrigues Telles Almeida ¹, Eliezer Fernando Pérez Pérez ^{1, 2, ✉}, Adriana Alicia Quiroga Carapia ³, Jesús Geovani Alcázar Sánchez ¹, Lorena Soto-Pinto ¹

1. *Department of Agriculture, Society and Environment, The College of the Southern Border (ECOSUR), San Cristóbal de las Casas, MEX*

2. *Ministry of Science, Humanities, Technology and Innovation, SECIHTI, Mexico, MEX*

3. *Outreach Coordination, The College of the Southern Border (ECOSUR), San Cristóbal de las Casas, MEX*

Received: June 12, 2026 | Review began: July 16, 2026 | Review ended: July 29, 2026 | Published: September 04, 2026

© Copyright 2026

This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Background

Qualitative social research faces a persistent methodological problem: how to move from lived narrative to analytical reconstruction without making that movement opaque. This is especially relevant in life-history research, where empirical material is dense, temporally layered, relational, and retrospectively organized. Existing Large Language Model (LLM)-assisted qualitative approaches have advanced tasks aligned with thematic analysis, such as coding, categorization, summarization, and pattern identification. However, they do not fully address life-history analysis, which requires reconstructing biographical temporality, narrative sequence, actor-event relations, and links between singular trajectories and broader social processes. Orchestrated Distant Reading (ODR) addresses this gap by organizing LLM assistance into a modular, staged, and researcher-verified workflow that makes each analytical transformation explicit.

Methods

This article introduces ODR, a generative pre-trained transformer (GPT)-based modular workflow for life-history analysis. The modules were selected according to the operations required to move from biographical narrative to traceable reconstruction: transcript cleaning, structured extraction, lexical and semantic mapping, construction of a project temporal baseline, interview temporal alignment, narrative profiling, cross-case comparison, inter-block synthesis, and controlled academic writing. During development, each module's instructional design was tested and refined to identify stable task definitions, input requirements, and output structures capable of producing satisfactory intermediate results under human review. The workflow was initially applied to two life-history interviews from a qualitative study on public interventions in coffee-growing territories in Chiapas, Mexico.

Results

The initial application showed that dense life-history interviews could be transformed into a traceable analytical sequence without collapsing interpretation into a single generative response. The workflow separated lexical recurrence, temporal reconstruction, actor-event relations, narrative profiling, and cross-case comparison. As procedural indicators, each module generated a predefined intermediate output linked to a specific input source, and each output was subject to a documented validation decision before reuse. This made it possible to distinguish biographical temporality from project-level historical chronology, identify mid-level comparative patterns, and preserve asymmetries between cases. Human verification remained central to preventing unsupported inference.

Conclusion

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. Cureus J AI Augment Res 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

ODR does not automate qualitative interpretation. Its contribution is to proceduralize analytical distance through a modular, auditable, and researcher-verified GPT workflow. The method offers a reproducible framework for artificial intelligence-augmented qualitative research with dense narrative materials that require interpretive depth and methodological transparency. Broader validation across additional datasets, qualitative methods, and research contexts is still required to evaluate the workflow's stability, adaptability, and limits.

Categories: Cross-Sector Human-AI Research Teams, Autonomous Research Pipelines, Methodological Innovation in AI Research

Keywords: orchestrated distant reading, gpt-assisted qualitative analysis, life-history research, analytical distance, ai-augmented research

Introduction

Social science research has always depended on a difficult movement: the transformation of lived experience into analytical reconstruction. This movement is not secondary to research; it is one of its constitutive operations. Social life is not received by the researcher as a transparent object. It appears through narrated memories, observed interactions, institutional documents, field notes, interviews, gestures, silences, classifications, and situated encounters. The task of social inquiry is therefore not simply to collect experience, but to reconstruct it under conditions that make interpretation communicable, contestable, and analytically meaningful. In this sense, the distance between experience and analysis is not a methodological defect. It is the very space in which scientific interpretation becomes possible. The problem is that this distance is often invoked more than it is procedurally documented.

This issue is central to the interpretive tradition in the social sciences. Weber's account of social science emphasized that social action becomes intelligible only insofar as the researcher reconstructs the meaning actors attach to their conduct [1]. Schutz [2] later radicalized this insight by showing that the social world is already organized through typifications, relevances, and intersubjective meanings before it becomes an object of scientific analysis. From this perspective, the researcher never encounters raw experience. What is encountered is already mediated by language, memory, expectation, social position, and practical orientation. The analytical problem is not how to eliminate mediation, but how to make mediation explicit enough that interpretation does not become arbitrary.

This is also why qualitative inquiry cannot be reduced to the extraction of themes from empirical material. Geertz's notion of thick description [3] established that the meaning of an action cannot be inferred from behavior alone; it depends on the symbolic and situational frame in which that action becomes socially intelligible. Goffman's interactional sociology [4] added another layer to this problem by showing that social encounters are performances in which people manage impressions, stage identities, and organize what becomes visible to others. Taken together, these perspectives suggest that qualitative materials are never transparent windows into subjectivity or social structure. They are situated productions that must be interpreted through controlled analytical operations.

The difficulty becomes particularly sharp in life-history research. Life histories allow researchers to examine large-scale social processes through concrete trajectories. Since the early work of Thomas and Znaniecki [5] on Polish migrants, biographical materials have been used to understand how historical transformations, migration, family relations, institutional changes, and cultural conflicts become sedimented in personal narratives. Ferrarotti [6,7] later argued that life histories should not be treated as merely illustrative documents, but as a specific form of sociological knowledge capable of articulating subjective experience and historical process. Their value lies precisely in the fact that they do not separate social structure from lived trajectory. They allow the researcher to observe how institutions, territories, crises, and forms of subjectivity are reconstructed within a narrated life.

However, the same feature that gives life-history methods their analytical power also produces a methodological risk. A narrated life tends to organize dispersed, discontinuous, and conflictive experiences into a sequence that appears meaningful after the fact. Bourdieu [8] described this risk as the "biographical illusion": the tendency to reconstruct a trajectory as if it were a coherent, linear, internally motivated story from beginning to end. His critique is not a rejection of

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

biography as empirical material, but a warning against treating retrospective narrative coherence as an objective property of the life itself. A trajectory must be reconstructed in relation to social positions, fields, institutions, historical constraints, and practical possibilities. Otherwise, the researcher risks converting narrative continuity into sociological explanation.

Clot [9] identified the inverse danger. If Bourdieu warns against the illusion of narrative coherence, Clot warns against the illusion that biography can be fully explained from the outside, as if subjective activity were only an effect of objective conditions. The subject does not invent a life freely, but neither does the subject merely reproduce structure. Life-history research therefore works in a difficult intermediate space: it must avoid romanticizing personal narrative, but it must also avoid erasing the activity through which people interpret, reorganize, and struggle over the meaning of their own experience. This tension defines one of the core methodological challenges of biographical research: how to move from lived narration to analytical reconstruction without either naturalizing the story or dissolving the narrator into structure.

The problem, then, is not whether the researcher should distance themselves from immediate experience. Analytical distance is unavoidable. The problem is how that distance is built, stabilized, documented, and justified. Classical qualitative methodology has addressed this issue through reflexivity, triangulation, credibility, dependability, confirmability, and audit trails [10,11]. Reflexivity requires the researcher to examine the conditions under which they produce interpretation, including their own position, assumptions, and relation to the field [12]. Triangulation seeks to avoid the overdependence on a single source, method, observer, or theoretical frame [10]. Audit trails make the research process reviewable by documenting how data, categories, decisions, and interpretations were produced [11]. Reporting standards such as COREQ and SRQR later translated these principles into formal expectations for qualitative research, asking researchers to report design, sampling, context, reflexivity, analytical procedures, and the derivation of findings from empirical material [13,14].

Yet a practical gap remains. Although qualitative methodology has long insisted on transparency, the actual movement from interview transcript to analytical claim is frequently difficult to inspect. This is especially true in dense qualitative materials such as life histories, where relevant evidence is distributed across long narrative sequences, temporal shifts, relational references, semantic recurrences, silences, contradictions, and retrospective reinterpretations. The researcher reads, marks, associates, compares, abstracts, returns to the material, changes categories, revises interpretations, and eventually writes findings. But the intermediate operations that connect raw narrative material to final analytical reconstruction are often weakly documented. The result is not necessarily poor research, but an epistemic opacity built into the craft of interpretation. The reader may see the transcript excerpts and the final argument, but not the full chain of transformations that made the argument possible.

Computational linguistics and computational text analysis emerged as partial responses to this problem of scale, consistency, and formalization. By treating text as data, computational approaches allowed researchers to observe distributions, frequencies, co-occurrences, semantic patterns, topic structures, and classificatory regularities across corpora too large to be handled by close reading alone [15,16]. In the social sciences, this made it possible to analyze political discourse, media corpora, policy documents, social media, interviews, and other textual materials through reproducible operations. These approaches did not eliminate interpretation. Rather, they shifted part of the interpretive process into explicit procedures: tokenization, cleaning, feature extraction, classification, clustering, topic modeling, dictionary construction, supervised annotation, and validation. Their promise was not that machines could understand social meaning in the human sense, but that certain textual regularities could be made visible, comparable, and testable.

Grimmer and Stewart's formulation remains important because it avoids technological naïveté. Automatic text analysis can expand the scope of inquiry, but it does not remove the need for close reading, substantive knowledge, validation, and theoretically informed interpretation [15]. Text-as-data approaches are powerful precisely when their limits are recognized. They can identify patterns, but the social meaning of those patterns depends on research design, contextual knowledge, and validation. A frequent word is not automatically a concept. A topic is not automatically a social process. A

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

co-occurrence is not automatically a relation of causality or meaning. Computational analysis therefore introduces a disciplined distance from the text, but it does not resolve the epistemological problem of interpretation. It provides formal representations that must be reinserted into a substantive analytical frame.

The tradition of distant reading sharpened this problem in the humanities. Moretti proposed distant reading as a way to observe literary systems through patterns that cannot be seen by reading a small number of canonical texts closely [17]. Jockers [18] extended this logic through macroanalysis, arguing that computational methods can reveal large-scale structures in literary history that remain invisible to individual reading. The contribution of distant reading is methodological: it redefined distance as a condition for seeing certain kinds of structure. Instead of assuming that interpretive depth depends exclusively on proximity to the individual text, distant reading showed that some patterns only appear when the researcher changes scale.

However, distant reading also revealed the cost of scale. The more texts are transformed into aggregate patterns, the more the researcher risks losing the density of situated meaning. In literary studies, this risk appears as the possible flattening of aesthetic, historical, and formal specificity. In social science research, the risk is even sharper: the transformation of lived experience into formal pattern can detach analysis from the temporal, territorial, institutional, affective, and relational density that gives qualitative material its value. This tension is particularly important for life-history research. A life history is not only a bag of words, a set of topics, or a sequence of extractable entities. It is a narrated reconstruction of time, relation, conflict, memory, position, and practical orientation. Computational formalization can support its analysis, but only if the procedure preserves enough traces of the narrative and contextual structure to avoid reducing biography to lexical distribution.

The methodological status of analytical distance can also be clarified through Varela's neurophenomenological reading of phenomenology and cognitive science. Varela's argument is not that lived experience can be solved by adding another theoretical layer to cognitive science, but that experience requires a rigorous method and an explicit pragmatics for its exploration [19]. This point is relevant beyond the specific field of consciousness studies. It helps specify a broader methodological problem: experience cannot be treated either as raw immediacy or as something to be eliminated through external explanation. It must be transformed through disciplined operations that make it available for intersubjective examination. In Varela's formulation, phenomenological reduction is not theoretical reduction in the eliminative sense. It is a practical gesture of suspension: a bracketing of habitual assumptions, automatic interpretations, and pre-given ways of organizing the phenomenon [19]. Reduction therefore provides a model for understanding analytical distance not as detachment from experience, but as a disciplined interruption of immediate interpretive capture.

This is central for the present argument. In life-history research, the researcher is not only confronted with narrated experience, but also with the temptation to recognize too quickly a coherent trajectory, a familiar sociological type, or an expected institutional pattern. A procedural method is needed precisely because the movement from lived narrative to analytical reconstruction requires successive acts of suspension, stabilization, and inscription. Varela's sequence from reduction to invariants offers a useful methodological analogy. Reduction opens the field by suspending habitual assumptions; variation allows the phenomenon to be examined across possible configurations; invariants emerge when aspects of experience are translated into communicable descriptions that can be shared, compared, and validated [19]. In this sense, invariants are not private impressions. They are stabilized forms of description that make experience available for intersubjective work.

Large Language Models (LLMs) intensify both sides of this debate. On one hand, LLMs create new possibilities for working with qualitative material. They can assist in summarization, segmentation, preliminary coding, comparison, extraction of actors and events, temporal organization, narrative profiling, and the transformation of unstructured text into structured outputs. Recent studies suggest that LLMs can perform some text-annotation tasks with high efficiency and, in certain conditions, with performance comparable to or better than crowd workers [20]. Other work has explored

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

their use in qualitative data analysis, deductive coding, and query-based analytical strategies [21-23]. These contributions show that LLMs cannot be dismissed as irrelevant to qualitative inquiry. They are already entering research practice, and their capacity to reorganize textual material is methodologically significant.

On the other hand, LLMs can make the opacity of interpretation worse. Their outputs often appear coherent, fluent, and analytically plausible even when the inferential path is unclear. Bender et al. warned that LLMs can produce convincing language without grounded understanding, creating risks of false authority and unexamined reproduction of bias [24]. In qualitative research, the danger is not only hallucination in the narrow sense of fabricating information. The deeper danger is the compression of multiple analytical operations into a single generative response. A single prompt may ask the model to read, summarize, categorize, compare, infer, theorize, and write. The result may be readable, but the process is difficult to audit. It becomes unclear which parts of the output are grounded in the text, which derive from the model's statistical priors, which reflect the prompt design, and which depend on the researcher's interpretive framing. Park et al. [25] recommend supporting the reproducibility and traceability of LLM-based workflows by providing executable prompts or complete experimental scripts and by documenting prompt iterations, execution procedures, model versions, hyperparameters, and stochasticity settings. These findings point in the same direction: the use of LLMs in qualitative research requires procedure, documentation, and human verification.

This article proposes Orchestrated Distant Reading (ODR) as a response to this combined problem. The term does not refer to autonomous agents. That distinction is important. The method does not assume that artificial systems possess independent agency, sociological judgment, or interpretive authority. Instead, ODR refers to a modular LLM-assisted workflow composed of constrained instructional modules. Each module performs a specific analytical operation under explicit restrictions, produces an intermediate output, and remains subject to human review before the next operation begins. The method is "orchestrated" because the value lies not in any isolated model response, but in the sequencing, separation, documentation, and verification of analytical operations. In this sense, the method draws from Varela's phenomenological and cognitive-scientific insight that experience does not become knowledge through immediate introspection or external explanation alone, but through disciplined mediations that suspend habitual assumptions, stabilize what is observed, and translate it into communicable forms [19,26].

The methodological contribution of ODR is therefore not the automation of interpretation. Its contribution is the proceduralization of analytical distance. It decomposes the movement from lived narrative to analytical reconstruction into staged operations: transcription cleaning, structured extraction, lexical and semantic mapping, temporal reconstruction, actor-event identification, narrative profiling, cross-case comparison, synthesis, and academic writing under evidence constraints. Each stage transforms the material in a limited way. Each transformation leaves an intermediate trace. Each trace can be reviewed, corrected, accepted, or rejected by the researcher. The workflow is designed to slow down interpretation rather than accelerate it indiscriminately. Its purpose is to prevent the premature collapse of transcript, pattern, interpretation, and theory into one fluent but opaque output. This procedural logic can be understood, in a limited methodological sense, as a movement from reduction to invariants: reduction as the suspension of immediate interpretive capture, and invariants as the provisional inscriptions that make aspects of the material available for comparison, review, and intersubjective validation [19].

In this sense, ODR inherits from distant reading the idea that distance can reveal patterns that close reading alone may miss. But it differs from classical distant reading in two ways. First, it is designed for dense qualitative material, including life-history interviews, rather than only for large-scale literary or documentary corpora. Second, it does not treat distance as a single change of scale. It treats distance as a sequence of controlled mediations. A life-history interview can be read as a narrative, a temporal structure, a lexical field, a network of actors, a set of territorial references, a sequence of institutional encounters, and a field of subjective elaboration. ODR does not claim that any one of these representations captures the life. Rather, it uses multiple partial representations to make the reconstruction of the life history more explicit and reviewable. These partial representations function as provisional analytical invariants within the research process: not as final truths about the life, but as stabilized forms of description that prevent interpretation from moving directly from immersion in the interview to final sociological claims.

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

ODR therefore extends distant reading rather than simply adapting it. Classical distant reading changes the scale of observation in order to reveal patterns that are not accessible through close reading alone [17,18]. The proposed workflow keeps this concern with distance, but relocates it from scale alone to procedural mediation. In dense qualitative materials, especially life histories, the key challenge is not only how to read more texts, but how to make the passage from lived narrative to analytical reconstruction explicit, staged, and reviewable. This also responds to recent empirical discussions of LLM-assisted qualitative analysis, which show the usefulness of LLMs for annotation, coding, and text organization while also emphasizing the need for validation, traceability, and researcher control [20-23]. The novelty of ODR lies in organizing these capacities into a sequential methodological architecture designed specifically for life-history analysis.

The method also draws from computational linguistics, but it does not reduce qualitative analysis to computational measurement. Frequencies, n-grams, co-occurrences, temporal markers, named entities, and semantic recurrences can help stabilize aspects of the corpus, especially when comparing long interviews or multiple cases. They can identify repeated lexical nuclei, associations between terms, shifts in thematic density, or recurrent links between actors and events. However, these indicators are not treated as findings in themselves. They are intermediate evidence. Their role is to discipline interpretation, not to replace it. A co-occurrence between “coffee” and “producers,” for example, may point to a relevant discursive association, but its meaning depends on narrative context, historical situation, institutional mediation, and the position of the speaker. ODR therefore uses computational representations as scaffolds for interpretation, not as substitutes for sociological analysis.

The method is especially relevant for life-history research because life histories require the researcher to move across several analytical levels at once. A biographical interview is temporal, but not simply chronological. It is subjective, but not merely individual. It is structured by social conditions, but not reducible to them. It contains memory, but memory is selective, relational, and reconstructed. It refers to institutions, but often through lived encounters rather than formal descriptions. It contains affective and evaluative orientations, but these may appear indirectly through emphasis, repetition, omission, or narrative rupture. A single-step LLM summary is poorly suited to this complexity because it tends to produce coherence. A modular workflow, by contrast, can separate temporal reconstruction from lexical mapping, actor identification from narrative profiling, and cross-case comparison from final interpretation. This separation does not guarantee validity, but it makes analytical movement easier to inspect and turns the production of distance into a documented sequence rather than an implicit interpretive leap.

This emphasis does not imply that life-history analysis is superior to other qualitative methodologies, but that it poses a different analytical problem. Thematic analysis, grounded theory, and computer-assisted qualitative data analysis software (CAQDAS)-supported coding are often organized around categories, themes, codes, or iterative comparison. These procedures can be highly effective for identifying patterned meanings across qualitative material. Life-history research, however, requires the researcher to preserve the internal movement of a trajectory: temporal sequence, narrative discontinuity, actor-event relations, institutional encounters, and the shifting connection between singular experience and broader social processes. For this reason, it particularly benefits from a workflow that can separate and then reconnect lexical, temporal, narrative, relational, and comparative layers without converting the biography into a list of themes or a single interpretive summary.

The first empirical application that grounds this article comes from a qualitative study of life histories related to public interventions in coffee-growing territories in Chiapas. In that preliminary work, two life-history interviews were processed through a modular workflow in order to examine how trajectories of professional formation, territorial mediation, agrarian memory, institutional participation, and coffee-related public interventions could be reconstructed through successive analytical layers [27]. The purpose of that initial study was not to claim definitive validation of the method. Its value lies elsewhere: it showed that a modular LLM-assisted sequence could maintain a traceable connection between interview transcripts, intermediate analytical representations, and final interpretive claims. In the context of life-history research, this is not a minor issue. It directly addresses the problem of how to move from narrated experience to analytical reconstruction without erasing the steps that make such reconstruction possible.

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

The case also clarifies why this method should not be framed as a generic artificial intelligence (AI) tool. ODR is not a software product, a chatbot technique, or an efficiency shortcut. It is a methodological architecture for controlling the relation between proximity and distance. Close reading remains necessary because the researcher must return to the interview, evaluate the plausibility of each output, correct distortions, and decide whether a pattern is sociologically meaningful. Distant reading becomes useful because it forces the material to pass through formal representations that can reveal patterns, associations, and structural features not immediately visible in immersive reading. Human verification remains central because neither computational linguistics nor LLMs can determine by themselves the historical, territorial, or theoretical significance of a life trajectory. The method depends on the interaction between these components.

The central argument of this article is that the use of LLMs in qualitative research should be judged less by the fluency of final outputs than by the quality of the procedural chain that produces them. A generative model can produce an elegant interpretation from a transcript, but elegance is not evidence. A model can summarize a life history, but summary is not analysis. A model can identify themes, but themes are not yet sociological findings. What matters is whether the research process can show how empirical material was cleaned, segmented, represented, compared, interpreted, verified, and written. ODR addresses this need by converting analytical distance into an explicit sequence of constrained operations. Its wager is simple: if qualitative research is always a reconstruction of experience, then AI-augmented qualitative research must make that reconstruction more visible, not less.

The article therefore contributes to three discussions at once. First, it contributes to qualitative methodology by proposing a way to document the passage from life-history narrative to analytical reconstruction through intermediate representations. Second, it contributes to computational social science and digital humanities by adapting the logic of distant reading to dense qualitative materials where scale is not the only problem and context cannot be treated as noise. Third, it contributes to current debates on AI-augmented research by offering a model of LLM use based on modularity, constraint, auditability, and human validation rather than automation or delegation. The broader claim is not that ODR solves the problem of interpretation. No method can do that. The claim is more modest and more defensible: it offers a structured way to make the construction of analytical distance more explicit, reproducible, and open to review.

Materials And Methods

Study design

This article presents a methodological study based on the design, application, and critical reconstruction of ODR, a modular LLM-assisted workflow for qualitative analysis of life-history interviews. The empirical material used to describe and test the method comes from an initial qualitative study on public interventions in coffee-growing territories in Chiapas, Mexico. That study examined two life-history interviews with participants whose trajectories were linked to institutional, academic, and territorial processes related to public programs in coffee-growing regions, specifically Sembrando Vida and PRONAI [27].

The purpose of the present article is not to evaluate the substantive effects of those programs. Rather, the two interviews are used as an empirical basis for reconstructing the methodological procedure through which life-history narratives were transformed into structured, traceable, and human-verified analytical outputs. The article therefore has a methodological and demonstrative design. It reconstructs how the workflow was applied, what kinds of intermediate representations it produced, and how those representations supported the movement from lived narrative to analytical reconstruction.

The study combines two methodological layers. The first is the collection and interpretation of life histories, grounded in the biographical tradition of the social sciences. The second is ODR, understood as a procedural strategy for analyzing dense qualitative material through constrained instructional modules, intermediate outputs, and human verification. The method is not presented as an autonomous AI system, nor as a replacement for qualitative interpretation. It is presented as a way to make the chain between narrative material, analytical transformation, and final interpretation more explicit.

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

Empirical material

The empirical material consisted of two life-history interviews conducted as part of a broader qualitative project on trajectories of life and public programs in coffee-growing territories of Chiapas. The interviews were selected because they occupied an exploratory and methodological role within the larger project. They were not treated as statistically representative cases. They were treated as historically dense trajectories through which it was possible to examine the relationship between professional formation, agrarian memory, territorial mediation, institutional participation, and public interventions in coffee-growing regions.

Both interviews were long-form narrative materials. They included biographical reconstruction, memories of rural and academic formation, references to institutions and public programs, descriptions of territorial work, and reflections on coffee-growing systems, agroforestry, and community engagement. For purposes of the present methodological article, the interviews were anonymized and identified only as ENT_01 and ENT_02. Personal identifiers were removed or generalized. The text refers to “interviewed persons” rather than named individuals in order to maintain the analytical focus on trajectories, positions, and processes rather than on personal identification.

The initial application of the method produced a methodological report in which each interview was processed through successive analytical layers: transcription cleaning, lexical extraction, co-occurrence analysis, temporal reconstruction, narrative profiling, cross-case comparison, and academic synthesis [27]. The present article reconstructs that workflow as a method in its own right.

Life-history data collection

The collection of the interviews was grounded in the life-history method, understood not as the extraction of biographical information from an individual subject, but as the construction of a situated narrative through an encounter between researcher and participant. This approach follows Ferrarotti’s conception of the biographical method as an autonomous form of sociological knowledge, rather than as an illustrative supplement to supposedly stronger quantitative findings [6,7]. In this view, a life history is not merely a personal testimony. It is a relational and interpretive event through which a society can be read through a biography.

Ferrarotti’s contribution is particularly relevant because it shifts the status of life histories from documentary material to interactive praxis. The interview is not a neutral channel through which an already complete life is transmitted. It is a social relation in which the researcher and the participant become implicated in the production of the narrative. This does not mean that the researcher invents the account, nor that the participant simply performs a role imposed by the researcher. It means that the life history emerges through trust, listening, situated questioning, and mutual orientation within a concrete research relationship. The biographical material is therefore not external to the conditions of its production; it is shaped by them.

This principle guided the treatment of the interviews. The purpose of the encounter was not to force the participants into a predetermined questionnaire structure, but to allow relevant problems, temporalities, and interpretive frames to emerge from the narrative itself. The interviewer introduced broad prompts concerning family background, education, territorial experience, professional formation, institutional participation, public programs, and coffee-growing territories. However, the sequence of the interview remained flexible. Participants were allowed to move between periods, themes, and memories according to the internal organization of their own account.

This decision followed the premise that life histories preserve the unpredictable dimension of social experience better than highly standardized instruments. Ferrarotti argues that standardized techniques tend to formalize subjects in advance by anticipating the possible answers they may give; life histories, by contrast, preserve a space for the problematic, the unexpected, and the not fully definable a priori [28,29]. This was methodologically important for the present study because the project was interested in trajectories of mediation, formation, and territorial engagement that could not be fully anticipated before the interview.

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

The method was also informed by Feixa's notion of autobiographical imagination. For Feixa, autobiographical imagination refers to the creative cooperation between a subject and a researcher in the construction of a writing on subjectivity. Its purpose is cognitive rather than merely literary: to read a social history through a life story [30]. This formulation was useful for clarifying the status of the interviews. The materials were neither treated as transparent expressions of individual consciousness nor as literary narratives detached from social analysis. They were treated as dialogically produced accounts through which broader historical, institutional, and territorial processes could become visible.

Feixa's formulation also emphasizes that life histories are structured by multiple dialogues: an internal dialogue between the narrator and their own memory, and an external dialogue between the narrator, the researcher, and the social world in which the narrative is produced [30]. This was especially important in the present case because the interviews did not only reconstruct individual trajectories. They also articulated relations among rural childhoods, academic institutions, agrarian histories, public policies, community processes, and coffee-growing territories. The empirical material therefore required a method capable of preserving both biographical singularity and social density.

Interview processing and preparation

The interviews were recorded, transcribed, anonymized, and prepared for analysis. The preparation of the material followed two principles. First, the transcript had to preserve the narrative organization of each life history. Second, it had to be transformed into a format suitable for systematic analysis without introducing distortions into the relative weight of lexical, semantic, and narrative patterns.

For this reason, all questions, prompts, and comments made by the interviewers were removed from the final analytical transcript. This decision was not merely editorial. It was necessary to prevent interviewer language from affecting later stages of the workflow, especially lexical frequency, n-gram extraction, co-occurrence analysis, semantic density, and the identification of recurrent narrative patterns. Since ODR depends on detecting distributions and associations within the interviewee's discourse, keeping interviewer interventions in the corpus could artificially inflate certain terms, introduce external thematic cues, or alter the proportional weight of categories that should emerge from the interviewed person's narrative.

The resulting corpus therefore consisted only of the discourse produced by each interviewed person. Elements of orality, hesitation, repetition, temporal uncertainty, affective emphasis, and narrative reformulation were preserved when they contributed to the meaning of the account. By contrast, interviewer turns were excluded, and purely mechanical transcription artifacts were corrected when they interfered with analytical processing. The objective was not to produce a polished literary version of the interview, but to generate a stable analytical corpus in which the patterns later identified could be attributed to the interviewed person's narrative rather than to the interactional scaffolding provided by the interviewer.

This preparation stage marked the transition from oral narrative to analyzable corpus. The life history remained a dialogically produced account, consistent with the biographical tradition that understands the interview as a situated encounter between researcher and participant. However, for the purposes of computational and comparative analysis, the final corpus had to isolate the narrative material produced by the interviewed person. This allowed the subsequent workflow to examine lexical recurrence, temporal organization, actor-event relations, and narrative structure without confusing the interviewee's discourse with the interviewer's prompts.

Orchestrated distant reading: general definition

ODR is defined here as a modular LLM-assisted workflow that decomposes qualitative analysis into a sequence of constrained, documented, and human-verified operations. The method is "orchestrated" because its analytical value does not reside in any single model response. It resides in the ordered relation among operations. Each module receives a specific input, performs a delimited task under explicit restrictions, produces a structured output, and passes that output to the next stage only after human review.

How to cite this article:

The term “module” is used deliberately. The workflow does not rely on autonomous agents. The instructional modules do not possess independent agency, sociological judgment, or authority over interpretation. They are configured analytical instruments. Their function is to stabilize specific transformations of the material: cleaning, structuring, temporalizing, comparing, or synthesizing. The researcher remains responsible for defining the task, reviewing the output, correcting inconsistencies, deciding whether the result is acceptable, and determining the analytical meaning of the material.

The workflow is based on four procedural principles. First, each module must operate only on the textual evidence provided to it. Second, each module must avoid filling gaps with external knowledge unless such knowledge is explicitly supplied as part of the input. Third, each module must produce an output in a stable and reviewable format. Fourth, no output is treated as final until it has been checked by the researcher. These principles are intended to reduce the risk of hallucination, false continuity, unsupported inference, and overinterpretation, which are known risks in generative language model use [24].

In methodological terms, ODR works by producing a controlled analytical distance from the interview. The researcher does not move directly from immersive reading to final interpretation. Instead, the material passes through a series of intermediate representations. These include cleaned transcripts, frequency tables, n-grams, co-occurrence structures, temporal sequences, actor-event maps, territorial references, narrative profiles, comparative matrices, and synthesized findings. These intermediate outputs function as provisional analytical inscriptions. They do not replace interpretation, but they make the interpretive chain more visible.

Analytical environment and instructional modules

The analytical workflow was implemented in ChatGPT through a set of public custom GPTs, each configured to perform a specific operation within the ODR sequence. OpenAI defines generative pre-trained transformers (GPTs) as custom versions of ChatGPT configured for a specific purpose, which can combine instructions, knowledge, and selected capabilities to create a tailored experience in ChatGPT [31]. OpenAI also specifies that GPTs can be accessed through direct links, the GPT Store, or workspace sharing settings, although users must be signed in and actual access may depend on plan limits, workspace permissions, and platform availability [32]. For this study, each instructional module was made publicly accessible through a direct GPT link so that other researchers could inspect, use, and replicate the workflow under comparable conditions.

Each public GPT functioned as a configured analytical environment: a version of ChatGPT constrained by a specific set of instructions, input requirements, output formats, and operational limits. In the present method, this configurability was used not to delegate analysis to an autonomous system, but to stabilize repeated analytical operations across interviews.

Each module corresponded to one delimited stage of the workflow: transcription cleaning, structured extraction, timeline generation, temporal alignment, narrative profiling, cross-case comparison, inter-block synthesis, and controlled academic writing. The modules were used sequentially. The output from one module became the input for the next only after human review. This sequencing is central to ODR: the method does not rely on a single prompt that asks the model to summarize, interpret, compare, and theorize at once. Instead, it separates these operations into discrete procedural moments, each one producing an intermediate output that can be inspected before the next transformation begins.

All public GPTs used in this study were configured with explicit methodological restrictions. They were instructed to work only with the evidence provided in the input, avoid filling gaps with external information, distinguish description from interpretation, mark uncertainty when evidence was insufficient, and produce outputs in stable formats suitable for subsequent processing. These restrictions were included to reduce the risk of unsupported inference, false continuity, and fluent but weakly grounded interpretation. The modules were therefore not designed to “discover” findings independently. Their function was to reorganize, stabilize, and make visible different dimensions of the corpus so that the researcher could evaluate them.

For reproducibility, the full configuration instructions used to build each public GPT are included in the Appendix. The Appendix reports the role assigned to each module, its task definition, input requirements, output structure, restrictions, and validation criteria. This is necessary because public access to a GPT link alone does not fully document the analytical

How to cite this article:

procedure. A public GPT allows other researchers to use the same configured module, but the methodological object that must be evaluated is the instruction set that governs the module's behavior. By providing the full instructions in the Appendix, the study allows the workflow to be replicated, modified, audited, or implemented in another LLM environment.

The public availability of the modules also served a second purpose: it made the workflow transferable beyond the original empirical case. Researchers working with other life-history corpora can use the same GPTs, adapt the instructions reported in the Appendix, or reconstruct equivalent modules in another platform. However, reproducibility is understood here in procedural rather than mechanical terms. Because LLM outputs can vary across model versions, system updates, and interaction contexts, exact verbal replication cannot be assumed. What can be replicated is the analytical sequence, the constraints imposed on each module, the type of intermediate output expected, and the human verification procedure applied at each stage.

Human review remained mandatory throughout the process. The researcher checked whether each module followed the assigned task, whether the output remained grounded in the transcript, whether unsupported information had been introduced, and whether the result was adequate for the next analytical phase. Outputs could be accepted, corrected, rerun, or discarded. The public GPTs therefore functioned as reproducible instructional environments, not as substitutes for methodological judgment. The analytical authority remained with the researcher, while the modules provided a standardized way to perform and document recurrent operations within the ODR workflow.

Phase 1: Transcription Cleaning and Corpus Stabilization

The first phase consisted of cleaning and normalizing the transcripts. The objective was to prepare the oral material for systematic analysis while preserving the narrative and semantic information necessary for life-history interpretation. This phase included lowercasing when required for computational operations, reducing non-analytical punctuation, standardizing obvious transcription inconsistencies, and limiting conversational noise that could distort lexical indicators.

The cleaning process did not aim to erase orality. It aimed to distinguish analytically relevant oral markers from elements that would interfere with downstream processing. Words and expressions linked to temporality, uncertainty, emphasis, affective orientation, institutional references, territorial markers, and narrative transitions were preserved. By contrast, purely mechanical repetitions or transcription artifacts could be reduced when they did not contribute to the meaning of the account.

When appropriate, the text was lemmatized to reduce morphological variation and improve consistency in lexical counts. The purpose of lemmatization was not to impose a fixed semantic system on the material, but to make recurring terms more visible across long interviews. This operation helped stabilize the transcript as a corpus. In methodological terms, this phase marked the transition from oral narrative to analyzable textual material.

Phase 2: Structured Extraction of Each Life History

The second phase transformed each cleaned transcript into a structured analytical template. This template included general corpus statistics, lexical frequency tables, n-grams, semantic density indicators, co-occurrence structures, narrative segmentation, biographical transitions, actor structures, actor-event mapping, territorial references, socio-environmental networks, and minimal indicators of narrative complexity.

This phase served two purposes. First, it imposed a common observational structure on heterogeneous interviews. Each life history remained singular, but each was processed through the same analytical dimensions. This made later comparison possible without forcing the interviews into a rigid coding scheme from the beginning. Second, it generated intermediate evidence that could be inspected before interpretation. Instead of moving directly from transcript to thematic conclusion, the researcher could examine how terms, actors, events, and narrative segments were distributed across the account.

How to cite this article:

Co-occurrence analysis was used to identify recurrent associations between terms within a defined textual window. This procedure drew on the principle that recurrent proximity between lexical items can indicate meaningful associations within a corpus [33]. In this workflow, co-occurrences were not treated as definitive meanings. They were treated as cues for interpretation. For example, a recurring association between terms related to “coffee,” “producers,” “community,” or “work” could indicate a discursive field that required closer qualitative examination. The researcher then returned to the transcript to determine whether the association corresponded to a substantive social relation, a narrative pattern, or a merely local linguistic recurrence.

The output of this phase was a structured profile of each interview. It converted the life history into a matrix of observable relations without reducing it to a closed coding system. The transcript remained available for close reading, while the structured extraction provided a map of potential analytical concentrations.

Phase 3: Timeline Generation and Temporal Alignment

The third phase of the workflow was divided into two sequential temporal operations. The first operation constructed a project-level temporal baseline. The second operation aligned the temporal structure of each life-history interview with that baseline. This distinction was methodologically important because the workflow did not treat historical chronology and biographical temporality as the same object. The project timeline provided a structured historical frame, while the temporal alignment module reconstructed how each interviewed person’s narrative moved through time.

The first temporal module was designed to transform contextual descriptions, historical events, institutional references, socio-environmental processes, and project-related temporal markers into a standardized project timeline. Its function was not to interpret history, but to produce a structured temporal baseline that could be used by the following alignment module. The module prioritized temporal structure over narrative explanation. It was instructed to work only with verifiable events, maintain traceability, avoid theoretical interpretation, and exclude events that lacked sufficient temporal reference.

Controlled external knowledge was allowed only in this first module and only for restricted purposes: to normalize dates of events explicitly mentioned in the input, to specify widely recognized historical periods, or to complete missing temporal information when the event itself was already present in the input. The module could not introduce unrelated events, expand the timeline with general historical information, infer events without direct textual basis, interpret causality, or mix scales without explicit connection to the project. For example, expressions such as “the 1990s,” “the late 1980s,” or “the early 2000s” were normalized into consistent temporal ranges when the reference was sufficiently clear. Events that could not be normalized with confidence were excluded.

The output of this first module followed a strict template. It included a chronological project baseline, the origin of each event, enriched events when date normalization or standard periodization was applied, excluded events with reasons for exclusion, temporal coverage, and event density by decade or relevant period. Events were classified as either directly present in the input or enriched through controlled temporal normalization. This output established the historical frame of the analysis and was used as the input for the second temporal module.

The second temporal module aligned each individual life-history interview with the project timeline. Its task was to identify temporalizable biographical events, assign them a temporal position, classify their temporal resolution, identify the type of temporal anchoring, estimate confidence, and relate them to the project baseline without modifying their original temporal placement. Unlike the first module, this module was not allowed to use external knowledge. It operated only with the project baseline and the structured interview template generated in the previous phases.

This distinction was central to the method. The project timeline could be enriched in a controlled way to improve historical standardization, but the life-history timeline could not be corrected from outside. Biographical time was treated as evidence in its own right. When an interviewee gave an explicit date, that date could not be modified, displaced, or adjusted to fit the project baseline. The module was explicitly prohibited from retroprojecting events, moving them to

How to cite this article:

earlier years, compressing the biography, or forcing alignment with historical events. The purpose was not to make the life story historically smoother, but to preserve the temporal logic of the narrative while making its relation to the project timeline inspectable.

The temporal alignment module used a hierarchy of evidence. Explicit dates had the highest priority. When explicit dates were absent, temporal location could be derived from age references, narrative sequence, or historical events mentioned directly in the interview. Historical-event anchoring was permitted only when the historical event appeared explicitly in the text. Each event was assigned a temporal resolution: exact date, range, decade, or relative position. Each event was also assigned an anchoring type: explicit, derived by age, derived by sequence, derived by historical event, or not resolvable. Finally, each temporal assignment received a confidence level: high, medium, or low.

The module also included rules for narrative continuity. Life histories rarely provide complete chronological sequences. They often include jumps, implicit transitions, long periods of activity, or present-oriented continuities that remain undated. For that reason, the temporal alignment module was required to cover the full narrative extension of the interview. When the narrative continued after the last explicitly dated event, the module could generate continuity events based on clear narrative progression. These events had to be marked as range or decade, anchored by sequence, and assigned medium or low confidence. This rule prevented the timeline from stopping artificially at the last explicit date while also preserving uncertainty.

The outputs of the alignment module included five structured sections: temporal alignment, project-baseline matching, temporal sequence, temporal lag, and temporal uncertainty. The temporal alignment section listed each event, its date or period, temporal resolution, anchoring type, and confidence level. The project-baseline matching section related biographical events to relevant project timeline milestones without altering the biographical date. The temporal sequence section reconstructed the order of events in the life history. The temporal lag section identified whether a biographical event was contemporary with, prior to, posterior to, or indeterminate in relation to a project-baseline event. The uncertainty section documented limitations such as lack of an explicit date, narrative ambiguity, relative reference, or conflicting information.

This two-step temporal phase allowed the workflow to distinguish between historical context and narrated time. The project baseline created a structured historical frame for the study. The alignment module then reconstructed each life history's temporal organization in relation to that frame, while preserving the interviewee's own chronology and uncertainty. This was especially important for life-history analysis because the aim was not to correct biography through history, but to examine how biographical trajectories relate to broader institutional, socio-environmental, and territorial processes.

The output from this phase was used in the following comparative module. By producing temporally aligned but non-corrected life histories, the workflow made it possible to compare trajectories according to sequence, timing, lag, simultaneity, and discontinuity. In this way, temporal analysis did not function as a background chronology. It became an intermediate analytical representation that helped connect individual trajectories with broader project-level processes while maintaining fidelity to the narrative evidence.

Phase 4: Narrative Profiling

The fourth phase produced a narrative profile for each interview. This phase moved from structured extraction to a controlled interpretive synthesis. The profile was organized around three dimensions: narrative location, linguistic characterization, and evidence-based interpretation.

Narrative location reconstructed the trajectory from early life to later institutional or professional consolidation. It identified spatial movements, educational transitions, professional shifts, political or institutional experiences, and changes in territorial engagement. This dimension remained close to the life-history structure and did not introduce theoretical abstraction beyond what was supported by the material.

How to cite this article:

Linguistic characterization used the outputs from the structured extraction phase to describe lexical nuclei, semantic concentrations, co-occurrence fields, and the organization of recurring terms. This made it possible to connect the narrative profile with corpus evidence. A trajectory was therefore not described only through remembered events, but also through the language that organized those events.

The third dimension, evidence-based interpretation, articulated narrative sequence and linguistic structure in order to identify controlled analytical patterns. These could include shifts in professional orientation, tensions between institutional and territorial forms of knowledge, recurrent links between scientific work and community engagement, or the progressive centrality of coffee-growing systems within the trajectory. However, the module was restricted from producing theoretical claims that could not be traced back to previous outputs.

This phase introduced a higher level of synthesis, but it did not return to unrestricted interpretation. The narrative profile was built from the transcript, the structured template, and the temporal reconstruction. In this sense, interpretation did not begin from the raw interview alone. It began from a life history already stabilized through several intermediate representations.

Phase 5: Cross-Case Comparison

The fifth phase compared the structured profiles of ENT_01 and ENT_02. The comparison was not based on free thematic interpretation. It was based on dimensions already produced in previous phases: lexical concentrations, narrative transitions, actor-event maps, temporal sequences, territorial references, and recurrent semantic fields.

The purpose of this phase was to move from singular trajectories to relational configurations without losing traceability. The module was instructed to identify only mid-level inferences grounded in observable recurrence, contrast, or sequence. It could not introduce broad theoretical generalizations unsupported by the material. For example, if both trajectories showed early rural socialization, later university formation, and subsequent territorial engagement, the module could identify this as a shared pattern. But it could not claim that this pattern represented a general generational structure unless the evidence supported such an inference.

The comparison produced analytical blocks. These blocks identified convergences, divergences, temporal asymmetries, shared actors or institutional fields, differences in the centrality of coffee, and recurrent links between knowledge, territory, and public intervention. The goal was not to erase the specificity of each case, but to identify where the two trajectories made visible a broader configuration of mediation between institutions, territories, and public programs.

Phase 6: Optional Inter-Block Synthesis

The sixth phase was designed for larger corpora and was only partially relevant in the initial two-interview application. Its purpose was to integrate multiple comparative blocks when the number of life histories increased. In a larger study, interviews can be processed in small groups, compared within blocks, and later synthesized across blocks.

The inter-block synthesis groups findings by controlled semantic similarity, counts recurrence across comparative blocks, identifies contradictions, detects gaps in evidence, and consolidates patterns that appear across multiple sets of cases. This phase is important because ODR is intended to scale beyond two interviews. Its architecture requires that each output be accumulable, verifiable, and compatible with subsequent synthesis.

Even at this stage, the method does not authorize unrestricted abstraction. The synthesis must indicate whether a finding is recurrent, exceptional, contradictory, or weakly supported. This allows the researcher to distinguish between robust patterns, emerging hypotheses, and interpretive possibilities requiring further evidence.

Phase 7: Controlled Academic Writing

The seventh phase converts the accumulated analytical outputs into academic prose. This phase is treated as part of the method, not as an external stage after analysis. The writing module receives structured outputs, comparative blocks, and human-selected interpretive priorities. Its central restriction is that it cannot introduce concepts, causal claims, or theoretical elaborations that are not grounded in the accumulated analytical chain.

How to cite this article:

The purpose of this phase is to preserve traceability between corpus, intermediate outputs, and final argument. Academic writing is not treated as the moment in which the researcher freely reinterprets the material from the beginning. It is treated as a controlled synthesis of already processed evidence. This does not remove the researcher's voice or judgment. It requires that final claims remain connected to the structured evidence produced in previous stages.

Human review is especially important in this phase. The researcher checks whether the written synthesis overstates the evidence, smooths out contradictions, introduces unsupported causality, or loses the specificity of the life histories. When necessary, the researcher returns to earlier outputs or to the original transcript. This recursive movement is central to the method. ODR is sequential, but it is not mechanically linear. Later stages can require correction of earlier outputs.

Figure 7 summarizes the full ODR workflow and clarifies the relationships between the different GPT-based instructional modules. It shows how the process moves from dialogical life-history interviews and transcript preparation to modular analytical operations, cross-case integration, and controlled academic writing. To support reproducibility and facilitate methodological evaluation, the Appendix provides an example of the complete workflow applied to a simulated life-history interview, including the intermediate outputs generated at each stage.

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

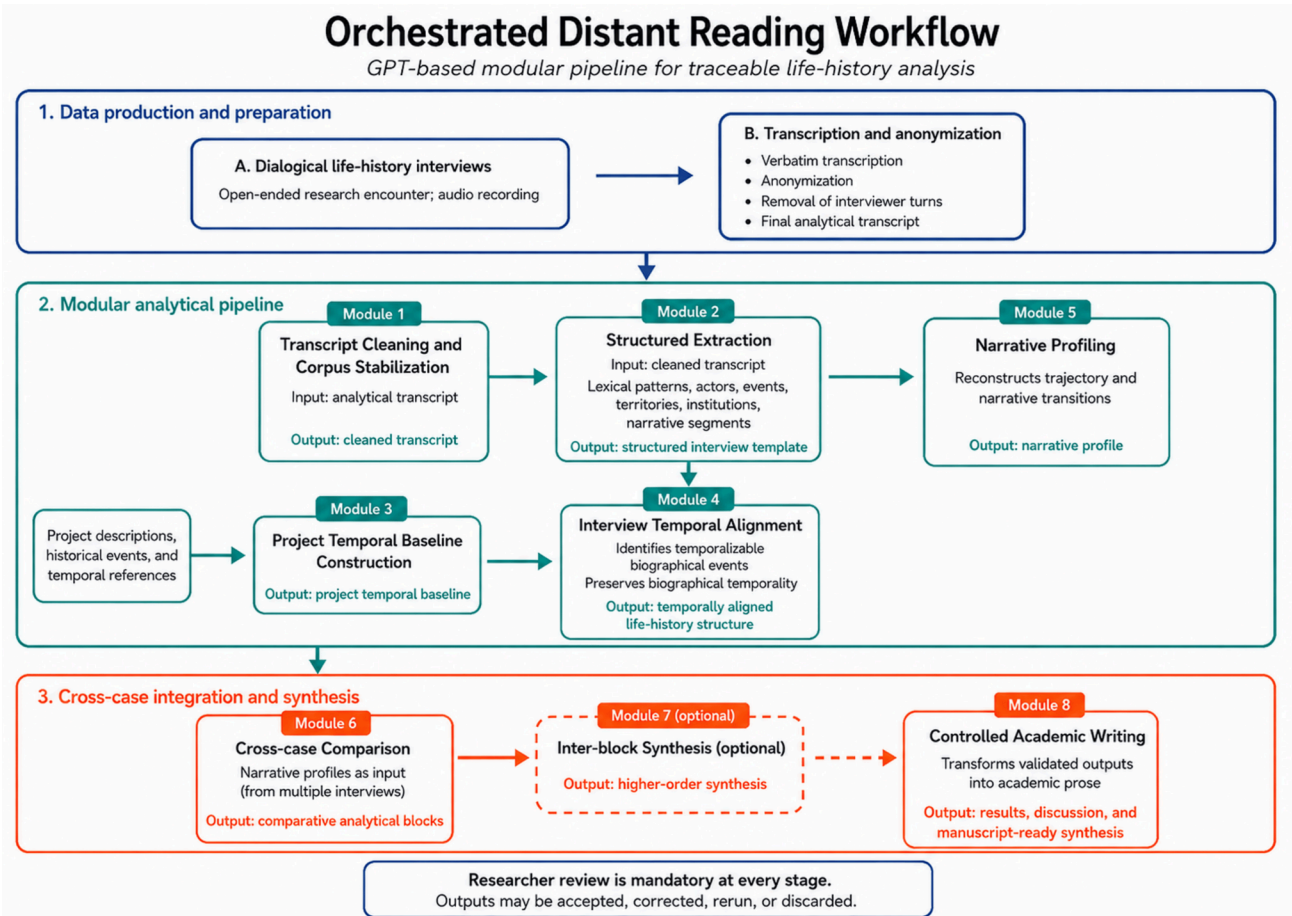


FIGURE 1: Workflow of Orchestrated Distant Reading

The figure presents the full sequence of the proposed GPT-based modular pipeline for traceable life-history analysis. It begins with data production and preparation, including dialogical life-history interviews, transcription, anonymization, and removal of interviewer turns. The modular analytical pipeline then proceeds through transcript cleaning and corpus stabilization, structured extraction, project temporal baseline construction, interview temporal alignment, and narrative profiling. These outputs feed the cross-case integration and synthesis stages, including cross-case comparison, optional inter-block synthesis, and controlled academic writing. The workflow emphasizes sequential dependency between modules and the requirement that outputs be reviewed by the researcher at each stage before being accepted, corrected, rerun, or discarded. GPT, Generative Pre-trained Transformer

Human verification procedures and decision criteria

Human verification was a mandatory procedure throughout the ODR workflow. No GPT-generated output was treated as valid by default, and no output was allowed to move to the next analytical stage without researcher review. The purpose of this verification process was not only to detect factual errors, but also to evaluate whether each output preserved the methodological function assigned to its module, remained grounded in the input material, and could be safely used as evidence for subsequent analysis.

Verification followed four general criteria. First, fidelity to the source material was assessed by checking whether the output remained anchored in the interview transcript, project timeline, or previous validated intermediate output. Outputs were examined for unsupported additions, invented details, unjustified reformulations, or distortions of the participant's narrative. Second, structural compliance was evaluated by verifying whether the module followed the required output format and included all expected fields. This was important because each output had to remain

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. Cureus J AI Augment Res 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

compatible with the next stage of the workflow. Third, uncertainty preservation was assessed by checking whether the module marked ambiguity, missing information, weak evidence, or low-confidence temporal reconstruction instead of converting uncertainty into apparent certainty. Fourth, analytical adequacy was evaluated by determining whether the output was suitable for downstream use without overstating its evidentiary status.

The decision-making process had four possible outcomes. An output was accepted when it was faithful to the input, structurally complete, methodologically consistent, and suitable for the next stage. An output was manually corrected when it contained minor problems, such as formatting inconsistencies, incomplete labels, or small classification errors that could be resolved without changing the substance of the analysis. An output was rerun when the module failed to follow instructions, omitted substantial information, produced an incomplete structure, or required a cleaner input formulation. An output was discarded when it introduced unsupported inferences, added information not present in the input, altered explicit dates, distorted the narrative sequence, or produced interpretations that could not be traced back to the source material.

Stage-specific verification criteria were also applied. During transcript cleaning, the researcher compared the cleaned transcript with the original transcription and, when necessary, with the original audio recording. This verification ensured that interviewer questions, prompts, and comments had been removed from the final analytical corpus while preserving the interviewed person's discourse. The researcher also checked that the cleaning process did not alter the meaning of the narrative, erase relevant temporal markers, or remove hesitations, repetitions, affective emphasis, or narrative reformulations that were analytically meaningful.

During structured extraction, the researcher checked whether the extracted lexical patterns, actors, events, territorial references, and narrative segments could be traced back to the transcript. No independent computational tool was used to recalculate or externally validate n-grams, frequencies, or co-occurrence structures. Instead, verification focused on textual traceability and thematic coherence: the researcher assessed whether the extracted patterns were consistent with the interviewed person's discourse and whether they corresponded to meaningful semantic, narrative, or relational configurations in the transcript. Items that could not be traced back to the interview, or that appeared thematically inconsistent with the source material, were removed, corrected, or the output was rerun.

During project temporal baseline construction, each event included in the baseline was reviewed to confirm that it was either directly present in the project input or had been enriched only through controlled temporal normalization or standard periodization. Events that were historically plausible but not directly connected to the project input were excluded. Date normalization was checked for consistency, and ambiguous temporal references were either marked as uncertain or omitted.

During interview temporal alignment, the researcher verified that explicit dates provided in the life-history material were not modified, displaced, or adjusted to fit the project timeline. Events derived by age, sequence, or historical reference were reviewed according to their assigned temporal resolution and confidence level. The researcher also checked whether the module preserved the distinction between biographical temporality and project-level historical chronology.

During narrative profiling, the researcher evaluated whether the profile preserved the internal logic of the life history, including discontinuities, tensions, shifts, and asymmetries. Interpretive statements were checked against the transcript, structured extraction, and temporal alignment. Profiles that converted the trajectory into an overly coherent or linear account were revised to preserve biographical complexity.

During cross-case comparison, the researcher verified that comparative claims were based on dimensions that had been produced in previous validated stages. Convergences and divergences were accepted only when they could be traced to lexical, temporal, narrative, territorial, or actor-event evidence. Claims that generalized beyond the cases analyzed or flattened differences between trajectories were revised or removed.

During inter-block synthesis, the researcher reviewed whether broader patterns were supported by recurrence across previously validated comparative outputs. This stage integrated cross-case findings into a broader synthesis while preserving distinctions between recurrent findings, divergent patterns, contradictions, and gaps in evidence. The

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

researcher checked that the synthesis did not smooth over inconsistencies, overgeneralize from weak evidence, or convert emerging patterns into definitive claims without sufficient support. When only one comparative block was available, this stage functioned as a controlled consolidation of the cross-case comparison rather than as a separate higher-order synthesis.

During controlled academic writing, the researcher reviewed whether the drafted prose introduced new theoretical claims, causal explanations, or empirical details that had not appeared in the validated analytical chain. Final claims were accepted only when they could be traced back to the transcript, an intermediate output, a validated comparison, or the inter-block synthesis. Text that was fluent but insufficiently grounded was revised or removed.

All verification decisions were treated as part of the audit trail of the study. The workflow therefore did not rely on GPT outputs as autonomous analytical judgments. Instead, GPT-generated outputs functioned as intermediate representations subject to human validation. This procedure ensured that the analytical chain remained inspectable and that final interpretations were based on researcher-verified evidence rather than on unreviewed generative output.

Reproducibility and documentation

The workflow was documented through a sequence of inputs, outputs, and researcher decisions. For each phase, the following elements were recorded: the input material supplied to the module, the instructional prompt or task definition, the expected output format, the generated output, the researcher's corrections or validation, and the decision to proceed, revise, or discard the output.

This documentation is necessary because LLM outputs can vary across model versions, prompt formulations, and interaction contexts. Reproducibility in this method does not mean that every run will produce identical wording. It means that the analytical procedure, restrictions, input-output sequence, and validation criteria are sufficiently explicit for another researcher to inspect, replicate, challenge, or adapt the workflow.

The study therefore treats reproducibility as procedural rather than purely mechanical. The objective is not to eliminate interpretation from qualitative analysis. The objective is to make interpretation more accountable by documenting the intermediate transformations that connect life-history material to analytical claims.

Ethical and anonymization procedures

Research Ethics, Informed Consent, and Confidentiality

The empirical material used to develop and test the ODR workflow consisted of life-history interviews with human participants. The use of this material was approved under the research protocol of the broader project from which the interviews were drawn. Participants were informed about the purpose of the study, the use of their interviews for qualitative analysis, the voluntary nature of participation, and the measures adopted to protect their identity and confidentiality. Informed consent was obtained before the interviews were used for analysis. No informed consent or confidentiality requirements were waived.

Because the empirical material consisted of life-history interviews, the workflow required strict anonymization. Names, direct institutional identifiers, and potentially identifiable personal details were removed or generalized before analytical processing. Participants were referred to as ENT_01 and ENT_02. The method also avoided unnecessary personalization by using neutral references to "interviewed persons" and by focusing on trajectories, positions, and processes rather than on biographical exposure.

The use of LLM-assisted processing introduced additional ethical concerns because the life-history accounts constituted sensitive qualitative material. Only anonymized or de-identified text was used in computational processing. Processing was conducted through a paid ChatGPT account configured so that submitted inputs and generated outputs were not used to train or improve the provider's models. In ChatGPT, this protection operates through the account's data controls: when the "Improve the model for everyone" option is disabled, conversations may remain available in the user's history but are excluded from model training. The researcher nevertheless retained responsibility for determining what material

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

was submitted to each module, what outputs were preserved, and what analytical claims could be made from them. No module was permitted to infer undisclosed private information, diagnose participants, or extend interpretations beyond the evidence contained in the de-identified material.

A further confidentiality advantage of the workflow is that the full primary interview material is not required at every stage of analysis. The original transcripts were used during transcript cleaning, verification, and the construction of the structured analytical template. The original audio recordings were never submitted to GPT-based processing and were used only by the researcher, outside the GPT environment, as a local verification source when checking transcript accuracy and cleaning decisions. Once the template, lexical patterns, temporal references, actor-event structures, territorial references, and other de-identified computational outputs had been generated and verified, subsequent stages relied primarily on these derived analytical representations rather than on the complete life-history transcripts. This reduced repeated exposure of the primary narrative material and allowed greater control over sensitive information.

The use of derived computational and analytical outputs does not eliminate ethical responsibility, but it helps preserve the source material by limiting the circulation of full transcripts after the initial stages of analysis. The workflow therefore combines prior project-level ethical approval, informed consent, anonymization, controlled use of GPT-based tools, staged reduction of primary data exposure, and researcher responsibility for all interpretive decisions. No identifiable participant information is reported in the manuscript.

This arrangement reduced, but did not eliminate, the ethical risks associated with processing sensitive qualitative data through a proprietary, externally hosted system. A limitation of the present method is therefore its dependence on the provider's infrastructure and data-governance policies. Future implementations should examine more controlled architectures, including locally hosted models or retrieval-augmented generation systems in which the research corpus remains within an institutionally governed repository and only the minimum information required for each analytical operation is retrieved. Such architectures are not inherently ethical, but, when accompanied by access controls, encryption, data minimization, retention limits, and auditable retrieval logs, they may offer stronger confidentiality, traceability, and institutional control over sensitive qualitative materials.

Methodological scope and limitations

The initial application of ODR was based on two life-history interviews. This number is sufficient for demonstrating the procedural architecture of the method, but not for claiming broad empirical generalization. The purpose of this article is therefore methodological rather than inferential. It shows how the workflow can be applied to dense qualitative material and how it can support traceability between narrative, intermediate representation, and interpretation.

The method also depends on the quality of the original interviews. ODR cannot compensate for poorly collected data, weak rapport, inadequate contextual knowledge, or superficial questioning. Because the life-history method is dialogical, the quality of the analytical corpus depends on the quality of the research encounter. The workflow begins after transcription, but its validity is shaped before transcription, in the relation between researcher and participant.

Finally, the method requires significant human labor. It is not an efficiency shortcut. The decomposition of analysis into multiple stages can slow down the research process. However, this is part of its methodological purpose. The workflow is designed to prevent premature synthesis, make interpretive movement inspectable, and preserve a documented chain between lived narrative and analytical reconstruction.

Results And Discussion

Transformation of life-history interviews into traceable analytical sequences

The initial application of ODR showed that dense life-history interviews could be transformed into a sequence of intermediate analytical outputs without collapsing the material into a single interpretive summary. In the two cases analyzed, the workflow produced cleaned transcripts, structured templates, lexical and semantic mappings, temporal

How to cite this article:

alignments, narrative profiles, and comparative blocks. The methodological result was not only a set of substantive findings, but a documented chain connecting interview discourse, intermediate representations, and final analytical claims [27].

This addressed one of the main weaknesses identified in the introduction: the opacity of the passage from lived narrative to sociological interpretation. Rather than moving directly from transcript to thematic synthesis, the workflow made each transformation inspectable. The researcher could verify how a lexical recurrence became a semantic field, how a temporal marker became part of a biographical sequence, and how a cross-case pattern was derived from comparable dimensions rather than from impressionistic reading.

Separation of analytical levels

The workflow made it possible to separate analytical levels that are often fused in conventional qualitative interpretation. In the two life histories, lexical patterns, temporal sequences, actor-event relations, institutional references, and narrative profiles were processed as distinct but connected outputs. This separation reduced the risk of treating narrative coherence as sufficient evidence.

For example, the analysis identified different lexical nuclei in each case. ENT_01 was organized around terms such as “land,” “biology,” “people,” “work,” and “field,” while ENT_02 showed stronger concentration around “coffee,” “work,” “community,” “producers,” and “field” [27]. These outputs did not function as findings by themselves. Their methodological value was that they provided a stable basis for returning to the interview and asking whether these recurrences corresponded to biographical transitions, institutional positions, or broader relational configurations.

The same occurred with co-occurrences and n-grams. Associations such as “biology-field,” “participatory diagnosis,” “participatory research,” “coffee-producers,” and “agroforestry systems” made visible recurrent links between technical knowledge, territorial work, and collective actors. The workflow therefore helped prevent an immediate interpretive leap from biography to theory. It first required the material to pass through observable linguistic and relational structures.

Reconstruction of temporality without forcing biographical coherence

The temporal phase produced one of the clearest methodological gains. The workflow distinguished between the historical baseline of the project and the internal temporality of each life history. This distinction allowed the analysis to identify long-term trajectories, periods of crisis, moments of professional reorientation, and recent reactivation of public-policy involvement without forcing the interviews into a single external chronology.

In the two cases, the temporal outputs showed that coffee did not operate simply as an initial biographical origin. Rather, it appeared as the result of accumulated trajectories involving rural socialization, university formation, institutional conflict, territorial engagement, and later participation in coffee-related public programs [27]. Methodologically, this result matters because it shows that temporal alignment can reveal lag, accumulation, and reorientation, instead of reducing the interview to a linear life story.

This directly responds to the risk of the biographical illusion. The workflow did not assume that a narrated life was already coherent. It reconstructed temporal relations, marked uncertainty, and allowed discontinuities or late-emerging centralities to remain analytically visible.

Cross-case comparison without flattening singular trajectories

The comparative stage showed that ODR can support cross-case analysis without reducing life histories to generic themes. The comparison identified a shared pattern described as “rural-university-territory”: both trajectories moved from early rural socialization to politicized university formation and then to prolonged technical-academic work in coffee-growing territories [27].

The methodological value of this pattern lies in how it was produced. It did not emerge from a single thematic coding operation. It was assembled across lexical evidence, temporal reconstruction, narrative profiling, and actor-event relations. The comparison therefore remained connected to the singularity of each case while making visible a broader

How to cite this article:

configuration.

At the same time, the workflow preserved asymmetry. Coffee appeared as a dominant semantic and operational axis in ENT_02, while in ENT_01 it became more relevant in later stages of the trajectory. This difference prevented the comparative output from flattening both interviews into the same narrative. The method made convergence and divergence visible at the same time.

Human verification as a condition of analytical validity

The application confirmed that human verification was not an auxiliary step, but a condition of analytical validity. The workflow generated structured outputs, but their inclusion in the next stage depended on researcher review. This was particularly important in three operations: deciding whether lexical recurrences were substantively meaningful, checking whether temporal alignments respected the interview's own chronology, and evaluating whether comparative patterns were supported by accumulated evidence.

The results therefore support a procedural understanding of AI-augmented qualitative analysis. The LLM-based modules contributed to stabilization, decomposition, and reorganization of the material. They did not determine the sociological significance of the findings. That significance depended on human review, contextual knowledge, and recursive movement between transcript, intermediate outputs, and final interpretation.

Methodological contribution of the initial application

Overall, the initial application demonstrated that ODR can address four methodological weaknesses in AI-assisted qualitative research. First, it reduces opacity by documenting the chain between transcript and interpretation. Second, it limits single-step generative overreach by separating cleaning, extraction, temporalization, comparison, and writing. Third, it preserves the density of life-history material by working across lexical, temporal, narrative, relational, and territorial dimensions. Fourth, it makes comparison possible without treating qualitative evidence as either purely subjective testimony or merely computational data.

The findings should not be read as full validation of the method. The empirical base consisted of two life-history interviews, and the application remains exploratory. However, the results show that the workflow can produce a reviewable methodological chain in which analytical distance is not assumed but constructed through successive, constrained, and human-verified operations. This is the central methodological finding of the study.

This contribution should not be understood as invalidating established qualitative approaches such as thematic analysis, narrative analysis, grounded theory, or CAQDAS-supported coding. Rather, ODR responds to a transversal problem shared by qualitative research: the need to construct analytical distance from lived or narrated experience without making that movement opaque. This problem becomes particularly acute in life-history research, where the unit of analysis is not only a theme or category, but a temporally organized trajectory through which social processes are reconstructed from a singular biographical account. In this sense, the workflow differs from other approaches not because it replaces their interpretive logic, but because it uses standardized computational procedures and GPT-assisted modules to support specific components of analysis - cleaning, extraction, temporal alignment, comparison, and synthesis - while preserving the interpretive responsibility of the researcher. Its contribution is therefore methodological: it shows how LLM assistance can be organized to make analytical distancing more explicit, staged, and reviewable, rather than reducing qualitative interpretation to automated coding or generative summarization.

As a synthesis of these methodological results, Table 1 presents the operational criteria through which ODR organizes LLM assistance into a controlled analytical workflow. For each module, the table identifies the specific analytical function performed, the restriction that prevents the module from exceeding its evidentiary scope, the validation criterion used by the researcher to determine whether the output can proceed, and the type of intermediate output produced. The table therefore summarizes the central result of the proposed method: analytical distance is not generated by a single automated interpretation, but by a sequence of constrained transformations whose outputs remain inspectable, revisable, and traceable to the preceding stages of analysis.

How to cite this article:

Module	Operational role	Main restriction	Validation criterion	Representative output unit
Module 1. Transcript Cleaning and Corpus Stabilization	Stabilizes the interviewee-only transcript for subsequent processing.	Must not remove narrative, temporal, affective, or territorial markers relevant to interpretation.	The cleaned transcript preserves the meaning and narrative structure of the original interviewee discourse.	Cleaned analytical transcript: interviewer turns removed; interviewee-only narrative retained.
Module 2. Structured Extraction	Converts the cleaned transcript into a structured analytical template.	Must not introduce categories or information not grounded in the transcript.	Extracted lexical, actor-event, territorial, and narrative elements can be traced back to the transcript.	CORPUS_STATISTICS; LEXICAL_FREQUENCY_TOP; COOCCURRENCE_STRUCTURE; ACTOR_EVENT_MAPPING; NARRATIVE_COMPLEXITY
Module 3. Project Temporal Baseline Construction	Builds the project-level historical and institutional timeline.	May use external knowledge only for controlled date normalization or standard periodization.	Included events are temporally anchored, project-relevant, and classified by origin.	PROJECT_BASELINE; EVENT_ORIGIN; ENRICHED_EVENTS; EXCLUDED_EVENTS; TEMPORAL_COVERAGE; EVENT_DENSITY
Module 4. Interview Temporal Alignment	Aligns biographical events with the project temporal baseline.	Must not modify explicit biographical dates or force events to fit the baseline.	Temporal assignments preserve biographical temporality and indicate resolution, anchoring type, and confidence.	TEMPORAL_ALIGNMENT: event
Module 5. Narrative Profiling	Produces a case-level narrative profile from the structured interview template.	Must not turn lexical or structural patterns into unsupported theoretical claims.	The profile remains grounded in the structured extraction and preserves narrative transitions and asymmetries.	Narrative location; lexical-semantic organization; actor-event interpretation; trajectory synthesis
Module 6. Cross-case Comparison	Compares narrative profiles across shared analytical dimensions.	Must not flatten singular trajectories into generic themes.	Comparative blocks identify convergences, divergences, and asymmetries supported by prior module outputs.	CONVERGENCES; DIVERGENCES; ASYMMETRIES; SHARED_PATTERN; CASE-SPECIFIC CONFIGURATION

How to cite this article:

Module 7. Inter-block Synthesis	Integrates comparative blocks when the corpus expands.	Must distinguish recurrence from weak, exceptional, or contradictory evidence.	Higher-order patterns remain linked to the comparative blocks from which they derive.	RECURRENT_PATTERN; EVIDENCE_STRENGTH; SUPPORTING_BLOCKS; LIMITATIONS
Module 8. Controlled Academic Writing	Converts validated outputs into academic prose.	Must not introduce unsupported causal claims, theoretical elaborations, or evidence not present in prior outputs.	Final writing remains traceable to validated intermediate outputs.	Evidence-constrained results text; discussion synthesis; methodological interpretation

TABLE 1: Operational role, constraints, validation criteria, and representative output units of the GPT-based modules

This table complements Figure 1 by detailing the operational role, main restriction, validation criterion, and representative output unit of each GPT-based module. The table emphasizes that the workflow is defined not only by the sequence of modules, but also by the constraints and review criteria that regulate each analytical transformation. Full demonstrative examples based on simulated life-history material are provided in Appendix

Limitations related to model dependence and output variability

The initial application also made visible several limitations of the proposed approach. ODR depends on the behavior of the LLM environment in which the public GPT modules are implemented. Although the modules were configured with explicit instructions, structured output formats, and evidence restrictions, their outputs remain model-dependent. Changes in the underlying LLM, platform updates, system-level behavior, or model availability may alter the wording, granularity, or classification choices produced by a given module.

The workflow is also sensitive to prompt formulation. Small differences in task instructions, input formatting, section labels, or the amount of contextual information supplied to a module may affect the resulting output. This sensitivity was partially mitigated by using stable public GPT modules, strict templates, and repeated human verification, but it was not eliminated. For this reason, the method should not be understood as producing mechanically identical outputs across all executions.

Reproducibility in ODR is therefore procedural rather than verbatim. The workflow makes it possible to reproduce the analytical sequence, the module functions, the input-output dependencies, the expected output structures, and the validation criteria. However, it does not guarantee that another execution, model version, or LLM platform will generate exactly the same text. The relevant standard of reproducibility is whether another researcher can inspect, rerun, compare, and evaluate the same procedural chain, not whether every generated sentence is identical.

Output variability is especially relevant for lexical and semantic extraction. In this initial application, the researcher did not use an independent computational tool to recalculate n-grams, frequency distributions, or co-occurrence structures. Verification focused on textual traceability and thematic coherence rather than on statistical replication of the extracted patterns. This means that the lexical outputs should be interpreted as intermediate analytical supports, not as independently validated computational measurements.

The integration of LLMs into qualitative research also requires attention to bias, hallucination, reliability, and epistemological status. LLM-assisted analysis does not occur in a neutral space. Model outputs may reproduce biases embedded in training data, dominant academic language, institutional vocabularies, gendered or racialized assumptions, and the linguistic hierarchies through which some forms of knowledge are more easily rendered as “analytical” than others. In the context of life-history research, this risk is especially important because biographical narratives often

include ambiguity, silence, contradiction, and situated meanings that should not be normalized into a coherent institutional or academic register. Bias may also enter through the researcher's own prompts, the selection of analytical categories, the organization of the workflow, and the decision to preserve or discard particular outputs.

Hallucination is therefore not only a factual problem, but also an interpretive one. In qualitative analysis, the most relevant risk is not simply that the model invents an event or a date, but that it fills narrative gaps, smooths tensions, imposes causal coherence, or translates uncertain biographical material into overly confident analytical language. For this reason, ODR treats GPT outputs as provisional intermediate representations rather than as findings. The workflow limits hallucination risk by constraining each module to a specific task, requiring outputs to remain traceable to the input material, preserving uncertainty markers, and subjecting each stage to human verification before the next step is performed.

Reliability must also be understood differently from deterministic computational reproducibility. Because GPT outputs may vary across runs, prompts, platforms, and model versions, the method cannot guarantee identical textual outputs. Its reliability depends instead on procedural stabilization: explicit module instructions, predefined output structures, documented input-output chains, researcher review, and an audit trail of accepted, corrected, rerun, or discarded outputs. This does not eliminate variability, but it makes variability inspectable and methodologically accountable.

The broader epistemological implication is that LLMs should not be understood as independent interpreters of qualitative material. Their role is better defined as the production of structured analytical mediations that can support, but not replace, human interpretation. ODR therefore shifts the epistemic status of AI assistance from authoritative interpretation to controlled inscription. The model helps generate intermediate forms - cleaned transcripts, lexical supports, temporal alignments, profiles, comparisons, and syntheses - but the validity of the analysis depends on whether these forms remain grounded in the material, ethically handled, and critically assessed by the researcher. In this sense, the workflow does not resolve the epistemological tensions of AI-assisted qualitative research; it provides a way to make them visible, staged, and reviewable.

These limitations do not invalidate the workflow, but they define its appropriate scope. The method is strongest when used to structure, document, and review qualitative analysis; it is weaker if treated as a deterministic computational pipeline. Future work should test the workflow across different LLM versions, compare outputs across repeated runs, evaluate the stability of prompts, and incorporate independent computational checks for lexical and co-occurrence measures when those indicators are used as formal evidence.

In this context, Table 2 situates ODR within the broader landscape of software-assisted and LLM-assisted qualitative analysis. The comparison is not intended to rank these approaches or to present ODR as a replacement for established qualitative methods. Rather, it clarifies how different analytical architectures distribute interpretation, software mediation, human oversight, and traceability. Conventional LLM prompting tends to concentrate several operations-reading, summarizing, coding, comparing, and writing-within a single generative exchange. Existing LLM-assisted qualitative frameworks have made important advances in coding, annotation, categorization, and summarization, but they are usually less oriented toward the specific demands of life-history reconstruction. CAQDAS platforms, in turn, offer powerful infrastructures for coding, retrieval, annotation, and data management, while leaving interpretive transformation largely in the hands of the researcher. ODR occupies a different position: it uses GPT-based modules to produce provisional analytical mediations, but it constrains those mediations through staged outputs, predefined functions, validation decisions, and repeated researcher review. The table therefore reinforces the central argument of this section: the risks of model dependence, hallucination, and variability cannot be eliminated, but they can be made more visible and methodologically accountable when LLM assistance is organized as a traceable chain rather than as a single act of automated interpretation.

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

Dimension	Conventional LLM prompting	Existing LLM-assisted qualitative analysis	Computer-assisted qualitative data analysis methods	Orchestrated Distant Reading
Representative approaches	Direct use of general-purpose conversational systems such as ChatGPT, Claude, or Gemini, without a predefined analytical architecture.	Includes LLM-in-the-loop [34], CollabCoder [35], the conversational AI-assisted qualitative analysis framework proposed by Friese [36], and TAMA [37].	Software environments such as NVivo, ATLAS.ti, MAXQDA, and similar platforms used to implement different qualitative methodologies.	A modular workflow specifically developed for the analysis of temporally layered life-history interviews using coordinated GPT-based analytical modules.
Primary analytical logic	A general-purpose model is asked to summarize, code, interpret, or compare qualitative material through one or a small number of prompts.	LLMs are integrated into defined stages of qualitative analysis, especially inductive or deductive coding, collaborative codebook development, theme generation, evidence retrieval, and dialogic interpretation.	Researchers use software to organize, code, retrieve, annotate, and compare qualitative materials according to a chosen methodological approach.	Analysis is divided into a coordinated sequence of specialized operations, each producing an explicit intermediate representation used by subsequent modules.
Typical workflow	Usually ad hoc, iterative, and dependent on the individual researcher's prompting strategy.	Usually follows a structured human–LLM interaction in which the model proposes codes, categories, themes, interpretations, or relevant passages and researchers accept, reject, discuss, or revise them. Approaches may also employ collaborative interfaces, conversational retrieval, or multiple specialized agents.	Usually organized around importing data, constructing a code system, coding excerpts, retrieving coded segments, memo writing, and producing queries or visualizations.	Follows a predefined modular sequence: transcript preparation, structured extraction, temporal baseline construction, temporal alignment, narrative profiling, cross-case comparison, synthesis, and controlled writing.
Principal unit of analysis	Often isolated excerpts, entire transcripts, or batches of text submitted within the model's context window.	Predominantly excerpts, coded segments, documents, codes, code groups, themes, or retrieved passages. Some approaches engage with larger bodies of text, but the complete biographical trajectory is not	Primarily coded excerpts, documents, cases, attributes, memos, and relationships established by the researcher.	Both the complete biographical trajectory and the structured analytical elements derived from it, including events, temporal positions, actors, relationships, transitions, and

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

		normally constituted as a distinct analytical unit.		retrospective reinterpretations.
Treatment of temporality	Temporal information may be summarized but is not ordinarily reconstructed through a systematic and comparable procedure.	Temporal references may be identified during coding, thematic interpretation, or conversational exploration, but temporality is generally treated as one possible analytical category rather than as a dedicated operation for reconstructing and aligning life trajectories.	Chronology can be coded, annotated, or visualized, but temporal reconstruction depends largely on manual researcher decisions and project-specific configurations.	Temporal analysis is a dedicated operation involving the construction of individual timelines, a shared temporal baseline, event alignment, and comparison between narrated trajectories.
Treatment of narrative context	Context may be compressed or lost when the model fragments material into codes, summaries, or decontextualized observations.	Iterative review, source retrieval, and human supervision can preserve part of the original context. Nevertheless, approaches centered on coding and theme generation continue to segment the material, while conversational approaches do not necessarily reconstruct the internal organization of complete narratives.	Source context can be preserved through links between codes and excerpts, although intensive coding may still fragment the narrative.	Intermediate outputs remain linked to the source material, while analytical modules alternate between localized extraction and reconstruction of the broader narrative trajectory.
Role of the computational system	The LLM may perform several analytical tasks simultaneously, often without clearly separating extraction, interpretation, comparison, and writing.	The LLM performs defined forms of analytical assistance, including suggesting codes, grouping labels, identifying patterns, generating themes, retrieving evidence, supporting analytical dialogue, or coordinating multiple analytical roles.	The software primarily supports data management, retrieval, coding, visualization, and documentation; interpretation remains principally with the researcher.	Different GPT-based modules perform delimited and sequentially dependent tasks under explicit instructions, while the researcher supervises transitions, validates outputs, and determines interpretive relevance.

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

Human oversight	Usually concentrated in prompt design and final review of the generated response.	Commonly organized through human-in-the-loop cycles in which researchers review, modify, accept, or reject model suggestions. The intensity of this oversight varies from continuous collaborative refinement to multi-agent analysis followed by specialist validation.	Continuous throughout coding, memo writing, category development, interpretation, and query construction.	Distributed across the workflow through verification checkpoints between analytical stages, especially when source material is transformed, aligned, compared, or incorporated into interpretive writing.
Traceability	Frequently limited because the final response may not reveal how particular interpretations were derived from specific passages or earlier model operations.	Usually stronger than in conventional prompting because codes, decisions, revisions, interaction histories, or links to supporting passages may be recorded. However, traceability is generally concentrated on relationships between excerpts, codes, themes, and interpretations rather than on a complete chain of narrative and temporal transformations.	Generally strong at the level of links between codes, excerpts, documents, and memos, although the reasoning connecting these elements may remain implicit.	Designed to preserve links between source passages, structured extractions, temporal representations, narrative profiles, comparative outputs, syntheses, and final claims.
Reproducibility	Often low because results are sensitive to prompt wording, model version, conversational history, and undocumented iteration.	Improved through structured workflows, prompt templates, codebooks, interfaces, decision records, or interaction histories. Nevertheless, reproducibility remains affected by model versions, parameters, implementation choices, and differences in the degree of human intervention.	Supported through documented codebooks, project files, coding protocols, memos, and query procedures, but dependent on researcher consistency.	Supported through standardized prompts, defined module functions, retained intermediate outputs, model and configuration documentation, and explicit human validation procedures.
Cross-case comparison	Commonly requested directly from the model,	Generally conducted by comparing codes,	Conducted through code matrices, case	Conducted only after individual narratives

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

	which may produce a synthetic comparison without exposing the transformations that support it.	themes, categories, retrieved passages, or model-generated syntheses across documents or participants. It does not ordinarily require the prior reconstruction of each case as an internally organized temporal trajectory.	attributes, queries, co-occurrence analysis, and researcher interpretation.	have been reconstructed and aligned through common temporal and analytical structures.
Relationship between close and distant reading	The distinction between analytical scales is rarely formalized.	Some approaches enable movement between corpus-level codes, themes, or interpretations and their supporting excerpts, but this change of scale is generally a function of coding, retrieval, or dialogue rather than an explicit methodological principle.	Researchers may move between excerpts, documents, codes, and corpus-level patterns, but the movement is methodologically determined outside the software.	Analytical distance is treated as a controlled and reversible movement between individual passages, complete trajectories, structured representations, and corpus-level patterns.
Typical output	Summaries, lists of codes, themes, interpretations, or draft prose.	Suggested or validated codes, coded excerpts, collaborative codebooks, thematic structures, retrieved evidence, dialogic interpretations, and reports produced through human–LLM collaboration.	Codebooks, coded corpora, thematic networks, matrices, memos, visualizations, and query results.	Auditable chains of source-linked extractions, timelines, temporal alignments, narrative profiles, comparative structures, syntheses, and controlled academic text.
Main limitation addressed by ODR	Lack of procedural differentiation, unstable outputs, limited auditability, and possible conflation of extraction with interpretation.	Existing frameworks substantially improve the structure and transparency of LLM-assisted qualitative analysis, but remain predominantly centered on coding, codebook development, thematic analysis, evidence retrieval, or conversational interpretation. They provide limited support	Strong support for organization and coding, but limited automation of complete, explicitly staged reconstruction of temporally layered life histories.	ODR does not eliminate interpretive judgment or claim superiority for every qualitative purpose. Its specific contribution is to make the sequence of computational transformations and human decisions involved in reconstructing and comparing life histories

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

		for the successive reconstruction, alignment, and comparison of complete, temporally layered biographical narratives.		more explicit, source-linked, and inspectable.
--	--	---	--	--

TABLE 2: Comparison of conventional LLM prompting, existing LLM-assisted qualitative analysis, computer-assisted qualitative data analysis, and Orchestrated Distant Reading

The table describes typical configurations rather than fixed properties. The LLM-assisted approaches grouped in the second comparative category differ substantially in their degree of automation, analytical assumptions, interface design, and distribution of agency between researchers and computational systems. LLM, Large Language Model; ODR, Orchestrated Distant Reading

Conclusions

This article introduced ODR as a methodological response to a persistent problem in qualitative social research: how to move from lived narrative to analytical reconstruction without making that movement opaque. The method was developed for dense life-history materials, where the value of the interview depends not only on what is said, but also on temporal sequence, narrative discontinuity, relational density, institutional mediation, and the way participants retrospectively organize their trajectories.

The central contribution of ODR is not the automation of qualitative interpretation. Its contribution is procedural. By decomposing analysis into constrained instructional modules, the workflow creates intermediate outputs that can be reviewed before becoming part of a final interpretation. Cleaning, structured extraction, lexical mapping, temporal alignment, narrative profiling, cross-case comparison, synthesis, and academic writing are treated as distinct operations rather than as a single generative act. This separation makes the analytical chain more visible and reduces the risk that fluent AI-generated prose will be mistaken for evidence.

The initial application to two life-history interviews showed that this workflow can support a traceable movement between interview transcripts, intermediate analytical representations, and final claims. The method made it possible to distinguish lexical recurrence from sociological interpretation, historical chronology from biographical temporality, and cross-case comparison from thematic flattening. In doing so, it addressed several weaknesses identified in current AI-assisted qualitative research: lack of procedural transparency, excessive dependence on single-step prompting, weak auditability, and insufficient control over the passage from empirical material to interpretation.

The approach also clarifies the role of computational linguistics and distant reading in qualitative inquiry. Frequencies, n-grams, co-occurrences, temporal markers, and actor-event structures are not treated as findings in themselves. They function as intermediate analytical inscriptions. Their purpose is to stabilize parts of the corpus, create points of return to the transcript, and support more disciplined interpretation. ODR therefore does not replace close reading. It reorganizes the relation between close reading and distant reading by making each movement between proximity and distance explicit.

The study has clear limitations. The empirical application was exploratory and based on two interviews. It does not establish generalizability across all forms of qualitative research, nor does it demonstrate that the workflow will perform equally well with other types of material, languages, or disciplinary contexts. The method also depends on the quality of the original interviews, the clarity of the instructional modules, and the researcher’s capacity to verify each output. It is not an efficiency shortcut. It may increase the time required for analysis because it obliges the researcher to document, inspect, and validate each step.

How to cite this article:

Even with these limits, the article argues that this additional procedural burden is methodologically justified. If AI is going to be used in qualitative research, the central issue is not only whether models can summarize, code, or classify text. The more important issue is whether researchers can show how AI-assisted operations participate in the production of evidence. ODR offers one possible answer: a modular, constrained, publicly documented, and human-verified workflow for making the construction of analytical distance more explicit.

Future work should test the method with larger life-history corpora, compare its outputs with conventional qualitative coding procedures, and evaluate its reproducibility across different LLM platforms and research teams. It should also examine how the workflow performs in multilingual, intercultural, and politically sensitive research contexts, where the risk of flattening experience into generic categories is especially high. The broader claim is therefore modest but important: ODR does not solve the problem of interpretation, but it provides a structured way to make interpretation more accountable when qualitative analysis is augmented by LLMs.

For replication and implementation, the workflow requires more than access to an LLM. Researchers should begin with a stable and well-documented corpus, explicit instructions for each GPT module, predefined output structures, and a clear procedure for human verification before each output is reused in the next stage. In this study, the GPTs used as instructional modules were made public and are reported as part of the method so that other researchers can inspect and apply the same analytical sequence. The appendices also provide an example of the expected prompts, outputs, and verification logic, allowing the workflow to be examined not only as a conceptual proposal but as an implementable methodological procedure. Recommended procedures include preserving the original transcript, documenting all cleaning decisions, separating extraction from interpretation, maintaining intermediate outputs as part of the audit trail, and recording whether each GPT-generated result was accepted, corrected, rerun, or discarded. When adapting the framework to other qualitative datasets, researchers should preserve its core logic - modularization, traceability, analytical separation, standardized computational assistance, and human verification - while adjusting the specific modules to the nature of the material. For example, datasets organized around focus groups, field notes, documents, or policy archives may require different units of segmentation, temporal anchors, or comparison criteria than life-history interviews. The workflow should therefore be replicated as a methodological sequence rather than as a fixed template: its value lies in making the movement from qualitative evidence to interpretation explicit, reviewable, and adaptable to the epistemic demands of each dataset.

Appendices

The Appendix provides two synthetic life-history interview excerpts created exclusively to illustrate the preprocessing stage of the Orchestrated Distant Reading workflow. These excerpts are not empirical data and were not used to produce the findings reported in the article. Their only purpose is to demonstrate how raw interview material is prepared before entering the GPT-based analytical modules. The outputs included in the Appendix are reproduced as generated by the GPT modules, with minimal formatting intervention. This decision was made deliberately so that readers can understand the structure, style, and level of standardization produced by the workflow itself, rather than a manually polished version of the output. When references appear at the end of an example output, they should not be read as part of the reference list of the present article. They are included because they form part of the standardized GPT response format used in the workflow, as a transparency mechanism for indicating the sources or evidentiary basis expected by the module.

The workflow was implemented in ChatGPT using GPT-4.4 through public custom GPTs configured for the analytical modules described in the article. The supplementary links to the ChatGPT conversations provide the actual interaction records used during the development and demonstration of the method, including prompt adjustments, errors, corrections, reruns, and author decisions during the use of the modules. These records were retained because they show the workflow as a practical process rather than as an idealized reconstruction. They also make visible how outputs were generated, reviewed, corrected, or rejected during implementation. Because the internal configuration of each public

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

GPT may need to be inspected in relation to platform access and version changes, the full instruction set for each GPT module is available from the authors upon reasonable request and will be gladly shared with researchers interested in replication or methodological evaluation.

PART 01

In both examples, the raw interview includes interviewer prompts and interviewee responses. The cleaned analytical transcript removes all interviewer questions, prompts, and comments. This preprocessing decision is necessary because later stages of the workflow analyze lexical frequencies, semantic patterns, temporal references, and narrative structures. Keeping interviewer interventions in the corpus could distort the relative weight of terms and patterns that should derive from the interviewee's discourse.

Synthetic Interview 1: SYN_01

Raw Interview Excerpt

Interviewer: Could you start by telling me where your relationship with coffee begins?

SYN_01: I was born in 1978 in a small community in the highlands. My parents had a small coffee plot. It was not large enough for the family to live comfortably, but it was enough for us to understand that coffee was not only a crop. It was work, family, debt, and a way of belonging to the community. As a child, I remember the harvest season as a time when everyone moved together: my parents, my siblings, neighbors, and sometimes relatives who came from nearby places.

Interviewer: Did coffee remain important when you left the community to study?

SYN_01: Not at first. When I left for secondary school in the early 1990s, I thought I was leaving coffee behind. I wanted to study because I did not want to repeat the same uncertainty that my parents lived with. But later, when coffee prices fell at the end of the 1990s, I began to understand that the problem was not only production. It was organization, market dependency, and the absence of institutions that really listened to small producers.

Interviewer: When did you begin working directly with producer organizations?

SYN_01: Around 2001 I joined a regional cooperative as a young technical assistant. At the beginning, my work was simple: visiting plots, writing notes, helping organize meetings, and translating technical language into something producers could use. Over time, I realized that technical assistance was never only technical. It was about trust. If people did not trust you, they would not change anything in the plot, even if the recommendation was correct.

Interviewer: Was there a moment that changed the direction of your work?

SYN_01: Between 2012 and 2014, the rust outbreak changed everything. Many families lost production, and some young people left the community. That period made me think more seriously about agroforestry systems. After 2016, I began participating in training processes on shade management, soil recovery, and diversification. Then, in 2019, when new public programs arrived in the region, I was invited to coordinate a small team that accompanied producers. That was when my previous experience with cooperatives, technical work, and community meetings became useful again.

Interviewer: How would you describe your role now?

SYN_01: I see it as mediation. I am not only a technician and not only someone from the community. I move between institutions, producers, reports, assemblies, and plots. Sometimes that position is uncomfortable because each side expects something different. But it is also where I think useful work can happen. Coffee taught me that territory is not a background. It is where relationships, memory, and institutions meet.

Cleaned Analytical Transcript

I was born in 1978 in a small community in the highlands. My parents had a small coffee plot. It was not large enough for the family to live comfortably, but it was enough for us to understand that coffee was not only a crop. It was work, family, debt, and a way of belonging to the community. As a child, I remember the harvest season as a time when everyone moved together: my parents, my siblings, neighbors, and sometimes relatives who came from nearby places.

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

Not at first. When I left for secondary school in the early 1990s, I thought I was leaving coffee behind. I wanted to study because I did not want to repeat the same uncertainty that my parents lived with. But later, when coffee prices fell at the end of the 1990s, I began to understand that the problem was not only production. It was organization, market dependency, and the absence of institutions that really listened to small producers.

Around 2001 I joined a regional cooperative as a young technical assistant. At the beginning, my work was simple: visiting plots, writing notes, helping organize meetings, and translating technical language into something producers could use. Over time, I realized that technical assistance was never only technical. It was about trust. If people did not trust you, they would not change anything in the plot, even if the recommendation was correct.

Between 2012 and 2014, the rust outbreak changed everything. Many families lost production, and some young people left the community. That period made me think more seriously about agroforestry systems. After 2016, I began participating in training processes on shade management, soil recovery, and diversification. Then, in 2019, when new public programs arrived in the region, I was invited to coordinate a small team that accompanied producers. That was when my previous experience with cooperatives, technical work, and community meetings became useful again.

I see it as mediation. I am not only a technician and not only someone from the community. I move between institutions, producers, reports, assemblies, and plots. Sometimes that position is uncomfortable because each side expects something different. But it is also where I think useful work can happen. Coffee taught me that territory is not a background. It is where relationships, memory, and institutions meet.

Synthetic Interview 2: SYN_02

Raw Interview Excerpt

Interviewer: Could you tell me how you became involved with coffee-growing communities?

SYN_02: My family did not produce coffee. I was born in 1985 in a municipal town where coffee was present everywhere, but not directly in my household. I knew coffee through neighbors, markets, trucks, and the people who came down from the communities during harvest. For me, coffee was part of the landscape before it became part of my work.

Interviewer: When did it become part of your work?

SYN_02: That happened when I entered university. Around 2006, I joined a student group that worked on rural development projects. At first, I thought the important thing was to bring technical information to communities. Later I understood that people already had knowledge, but institutions often did not recognize it. In 2010, I participated in a university extension project with coffee producers, and that changed my way of thinking.

Interviewer: What changed in your way of thinking?

SYN_02: I stopped seeing producers as people who needed to be instructed. I began to see them as people who were already experimenting, comparing soils, observing shade, and deciding what risks they could take. My role became less about explaining and more about listening, organizing information, and helping groups compare what they were already doing.

Interviewer: Did the rust outbreak affect that process?

SYN_02: Yes, strongly. Between 2012 and 2014, many families were worried because production dropped and some plants were lost. In those years, meetings became more intense. People talked about varieties, shade, pruning, soil, and whether it made sense to continue with coffee. For me, that period showed that a crisis is not only technical. It affects family decisions, migration, debt, and trust in institutions.

Interviewer: And what happened after that period?

SYN_02: After 2015, I continued working with producer groups in participatory diagnosis processes. We mapped problems, but also resources: springs, shade trees, family labor, local seed, and collective experience. Later, in 2019, when public programs expanded in the region, some organizations asked us to help systematize information. I was not

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

coordinating a program directly, but I worked with teams that needed to understand what was happening in the communities.

Interviewer: How do you understand your role today?

SYN_02: I think of it as facilitation. I do not come from a coffee-producing family, so I cannot speak as if I had lived that history from inside. But I have spent many years listening to producers and accompanying collective work. My task is to help connect what communities know with what institutions ask for, without reducing everything to forms, numbers, or reports. The most difficult part is that institutions want quick evidence, while community processes need time.

Cleaned Analytical Transcript

My family did not produce coffee. I was born in 1985 in a municipal town where coffee was present everywhere, but not directly in my household. I knew coffee through neighbors, markets, trucks, and the people who came down from the communities during harvest. For me, coffee was part of the landscape before it became part of my work.

That happened when I entered university. Around 2006, I joined a student group that worked on rural development projects. At first, I thought the important thing was to bring technical information to communities. Later I understood that people already had knowledge, but institutions often did not recognize it. In 2010, I participated in a university extension project with coffee producers, and that changed my way of thinking.

I stopped seeing producers as people who needed to be instructed. I began to see them as people who were already experimenting, comparing soils, observing shade, and deciding what risks they could take. My role became less about explaining and more about listening, organizing information, and helping groups compare what they were already doing.

Yes, strongly. Between 2012 and 2014, many families were worried because production dropped and some plants were lost. In those years, meetings became more intense. People talked about varieties, shade, pruning, soil, and whether it made sense to continue with coffee. For me, that period showed that a crisis is not only technical. It affects family decisions, migration, debt, and trust in institutions.

After 2015, I continued working with producer groups in participatory diagnosis processes. We mapped problems, but also resources: springs, shade trees, family labor, local seed, and collective experience. Later, in 2019, when public programs expanded in the region, some organizations asked us to help systematize information. I was not coordinating a program directly, but I worked with teams that needed to understand what was happening in the communities.

I think of it as facilitation. I do not come from a coffee-producing family, so I cannot speak as if I had lived that history from inside. But I have spent many years listening to producers and accompanying collective work. My task is to help connect what communities know with what institutions ask for, without reducing everything to forms, numbers, or reports. The most difficult part is that institutions want quick evidence, while community processes need time.

PART 02

The following material presents the representative output generated by the GPT-based analysis module after processing the two synthetic cleaned life-history transcripts. These examples are included for reproducibility and demonstration purposes only. They do not correspond to empirical interview data and were not used to produce the findings reported in the article. Their function is to illustrate the type of structured output produced by the analytical module before subsequent stages of temporal alignment, narrative profiling, cross-case comparison, and controlled synthesis.

GPT-Based Analysis Output:

CORPUS_STATISTICS

ID_INTERVIEW: ENT_1
TOTAL_WORDS: 389
TOTAL_LEXICAL_WORDS: 218
LEXICAL_DENSITY: 0.5604
LEXICAL_FREQUENCY_TOP
FORMAT:
term frequency relative_frequency
coffee 7 0.0321
family 5 0.0229
people 5 0.0229
work 5 0.0229
communities 4 0.0183
institutions 4 0.0183
producers 4 0.0183
information 3 0.0138
groups 3 0.0138
shade 3 0.0138
part 3 0.0138
NGRAMS
FORMAT:
ngram frequency
coffee-producing family 1
municipal town 1
rural development projects 1
technical information 1
university extension project 1
coffee producers 1
comparing soils 1
observing shade 1
family decisions 1
participatory diagnosis processes 1
producer groups 1
collective experience 1
public programs 1
systematize information 1
collective work 1
quick evidence 1
community processes 1
SEMANTIC_DENSITY

FORMAT:
core_term relative_weight
coffee 0.0321
family 0.0229
people 0.0229
work 0.0229
communities 0.0183
institutions 0.0183
producers 0.0183
information 0.0138
groups 0.0138
shade 0.0138
COOCCURRENCE_STRUCTURE
WINDOW: 5
FORMAT:
central_term → associated_term frequency
coffee → family 2
shade → soil 2
coffee → communities 1
coffee → producers 1
coffee → work 1
communities → institutions 1
information → communities 1
information → groups 1
producers → people 1
producers → soils 1
producers → shade 1
producer_groups → resources 1
resources → springs 1
resources → shade_trees 1
resources → family_labor 1
resources → local_seed 1
institutions → evidence 1
community_processes → time 1
NARRATIVE_SEGMENTATION
FORMAT:
segment text_range dominant_lexical_fields
1 paragraph 1 family, coffee, municipal town, neighbors, markets, trucks, communities, harvest, landscape, work
2 paragraph 2 university, student group, rural development, technical information, communities, knowledge,

institutions, extension project, producers

3 | paragraph 3 | producers, people, experimenting, soils, shade, risks, listening, information, groups

4 | paragraph 4 | families, production, plants, meetings, varieties, shade, pruning, soil, coffee, crisis, migration, debt, institutions

5 | paragraph 5 | producer groups, diagnosis processes, resources, springs, shade trees, family labor, local seed, collective experience, public programs, organizations, teams, communities

6 | paragraph 6 | facilitation, coffee-producing family, history, listening, producers, collective work, communities, institutions, forms, numbers, reports, evidence, community processes, time

BIOGRAPHICAL_TRANSITIONS

FORMAT:

origin_event → destination_event

born in 1985 in a municipal town → entered university

entered university → joined a student group around 2006

joined a student group → worked on rural development projects

bringing technical information to communities → understanding that people already had knowledge

participated in a university extension project in 2010 → changed way of thinking

seeing producers as people who needed instruction → seeing producers as people who were experimenting

explaining → listening, organizing information, and helping groups compare

production dropped and plants were lost between 2012 and 2014 → meetings became more intense

continued working with producer groups after 2015 → mapped problems and resources

public programs expanded in 2019 → organizations requested help systematizing information

listening to producers and accompanying collective work → helping connect community knowledge with institutional requests

ACTOR_STRUCTURE

FORMAT:

actor | type | frequency

family | collective | 5

neighbors | person | 1

people | person | 5

communities | collective | 4

university | institution | 2

student group | collective | 1

institutions | institution | 4

coffee producers | person | 1

producers | person | 3

families | collective | 1

producer groups | collective | 1

public programs | program | 1

organizations | institution | 1

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

teams | collective | 1
university extension project | program | 1
coffee | resource | 7
soils | resource | 1
shade | resource | 3
plants | resource | 1
varieties | resource | 1
soil | resource | 1
springs | resource | 1
shade trees | resource | 1
family labor | resource | 1
local seed | resource | 1
collective experience | resource | 1

ACTOR_EVENT_MAPPING

FORMAT:

event | actors
birth in a municipal town in 1985 | family
entry into university | university
joining a student group around 2006 | student group
rural development projects | student group, communities
university extension project in 2010 | university, coffee producers
experimenting, comparing soils, observing shade, and deciding risks | producers
production drop and plant loss between 2012 and 2014 | families
intensification of meetings | people
participatory diagnosis processes after 2015 | producer groups
mapping problems and resources | producer groups
expansion of public programs in 2019 | public programs
systematization of information | organizations, teams
listening and accompanying collective work | producers
connecting community knowledge with institutional requests | communities, institutions

TERRITORIAL_REFERENCES

FORMAT:

place | frequency
municipal town | 1
communities | 4
markets | 1
region | 1

SOCIO_ENVIRONMENTAL_NETWORK

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

FORMAT:
actor → resource → practice
producers → soils → comparing
producers → shade → observing
producers → risks → deciding
people → varieties → talking
people → shade → talking
people → pruning → talking
people → soil → talking
people → coffee → continuing
producer groups → springs → mapping
producer groups → shade trees → mapping
producer groups → family labor → mapping
producer groups → local seed → mapping
producer groups → collective experience → mapping
communities → knowledge → connecting
institutions → evidence → requesting
NARRATIVE_COMPLEXITY
FORMAT:
indicator value
number_of_actors 26
number_of_events 14
approximate_lexical_diversity 0.7844
semantic_density 0.1972
CORPUS_STATISTICS
ID_INTERVIEW: ENT_02
TOTAL_WORDS: 382
TOTAL_LEXICAL_WORDS: 213
LEXICAL_DENSITY: 0.5576
LEXICAL_FREQUENCY_TOP
FORMAT:
term frequency relative_frequency
coffee 5 0.0235
community 5 0.0235
technical 5 0.0235
work 5 0.0235
producer 4 0.0188

plot 4 0.0188
family 3 0.0141
institution 3 0.0141
parent 3 0.0141
meeting 2 0.0094
people 2 0.0094
production 2 0.0094
trust 2 0.0094
NGRAMS
FORMAT:
ngram frequency
small community 1
small coffee plot 1
harvest season 1
secondary school 1
coffee prices 1
market dependency 1
small producers 1
regional cooperative 1
technical assistant 1
technical language 1
technical assistance 1
rust outbreak 1
agroforestry systems 1
training processes 1
shade management 1
soil recovery 1
public programs 1
small team 1
community meetings 1
SEMANTIC_DENSITY
FORMAT:
core_term relative_weight
coffee 0.0235
community 0.0235
technical 0.0235
work 0.0235
producer 0.0188
plot 0.0188

family 0.0141
institution 0.0141
parent 0.0141
meeting 0.0094
people 0.0094
production 0.0094
trust 0.0094
COOCCURRENCE_STRUCTURE
WINDOW: 5
FORMAT:
central_term → associated_term frequency
coffee → plot 2
coffee → parent 2
coffee → family 1
coffee → community 1
coffee → price 1
coffee → territory 1
community → parent 1
community → sibling 1
community → people 1
community → meeting 1
producer → institution 1
producer → technical 2
producer → plot 1
technical → assistant 1
technical → language 1
technical → assistance 1
technical → work 1
trust → people 1
production → family 1
shade → soil 1
program → region 1
institution → producer 1
institution → relationship 1
NARRATIVE_SEGMENTATION
FORMAT:
segment text_range dominant_lexical_fields
1 paragraph 1 birth, community, highlands, parents, coffee plot, family, work, debt, harvest, siblings, neighbors, relatives

2 | paragraph 2 | secondary school, coffee, parents, uncertainty, prices, production, organization, market dependency, institutions, producers

3 | paragraph 3 | regional cooperative, technical assistant, plots, notes, meetings, technical language, producers, technical assistance, trust

4 | paragraph 4 | rust outbreak, families, production, young people, community, agroforestry systems, training, shade, soil, diversification, public programs, team, producers

5 | paragraph 5 | mediation, technician, community, institutions, producers, reports, assemblies, plots, work, coffee, territory, relationships, memory

BIOGRAPHICAL_TRANSITIONS

FORMAT:

origin_event → destination_event

born in 1978 in a small community → left for secondary school in the early 1990s

left for secondary school → thought coffee was being left behind

coffee prices fell at the end of the 1990s → began to understand organization, market dependency, and absence of institutions

joined a regional cooperative around 2001 → worked as a young technical assistant

visiting plots and organizing meetings → realized technical assistance was about trust

rust outbreak between 2012 and 2014 → families lost production and young people left

rust outbreak → thought more seriously about agroforestry systems

participated in training processes after 2016 → worked on shade management, soil recovery, and diversification

public programs arrived in 2019 → invited to coordinate a small team

previous cooperative and technical experience → became useful again

technical and community position → mediation between institutions, producers, reports, assemblies, and plots

ACTOR_STRUCTURE

FORMAT:

actor | type | frequency

parents | person | 3

siblings | person | 1

neighbors | person | 1

relatives | person | 1

small producers | person | 1

producers | person | 3

young technical assistant | person | 1

people | person | 2

families | collective | 1

young people | collective | 1

regional cooperative | institution | 1

cooperatives | institution | 1

institutions | institution | 3

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI* Augment Res 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

public programs | program | 1

small team | collective | 1

community | collective | 5

coffee | resource | 5

coffee plot | resource | 1

plots | resource | 2

production | resource | 2

shade | resource | 1

soil | resource | 1

territory | resource | 1

ACTOR_EVENT_MAPPING

FORMAT:

event | actors

birth in 1978 in a small community | parents, family, community

harvest season | parents, siblings, neighbors, relatives

departure for secondary school in the early 1990s | parents

coffee price fall at the end of the 1990s | small producers, institutions

joining a regional cooperative around 2001 | regional cooperative, young technical assistant

plot visits and meeting organization | producers

translation of technical language | producers

technical assistance based on trust | people, producers

rust outbreak between 2012 and 2014 | families, young people, community

participation in training after 2016 | producers

arrival of public programs in 2019 | public programs

coordination of a small team | small team, producers

mediation between positions | institutions, producers, community

TERRITORIAL_REFERENCES

FORMAT:

place | frequency

community | 5

highlands | 1

nearby places | 1

region | 1

territory | 1

SOCIO_ENVIRONMENTAL_NETWORK

FORMAT:

actor → resource → practice

parents → coffee plot → work

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

family	→	coffee	→	work
family	→	coffee	→	belonging
parents	→	coffee	→	production
small producers	→	coffee	→	production
young technical assistant	→	plots	→	visiting
producers	→	technical language	→	use
people	→	plot	→	change
families	→	production	→	lost
young people	→	community	→	left
producers	→	shade	→	management
producers	→	soil	→	recovery
small team	→	plots	→	accompanied
technician	→	plots	→	mediation
community	→	territory	→	relationships
NARRATIVE_COMPLEXITY				
FORMAT:				
indicator		value		
number_of_actors		23		
number_of_events		13		
approximate_lexical_diversity		0.7981		
semantic_density		0.2066		

Source link: <https://chatgpt.com/share/6a55531a-25a8-83e8-bd4d-818a972b00a3>

PART 03

Representative Output from the Profile Generator GPT

The following text presents a representative output generated by the Profile Generator GPT after processing the structured analysis of two synthetic life-history interviews. This output illustrates how the module transforms lexical, temporal, actor-based, and narrative evidence into a controlled analytical profile. The examples are synthetic and are included only to demonstrate the workflow.

Profile Generator Output:

ENT_02
The actor was born in the late 1970s in a small highland community, within a family connected to coffee cultivation. Their childhood unfolded through work on the family plot, collective participation during harvest periods, and everyday relations with relatives and neighbors. In the early 1990s, the actor moved to nearby localities to attend secondary school, an experience that introduced an initial separation from the community setting. Although the actor initially perceived coffee as something being left behind, the economic uncertainty

experienced by the family and the fall in coffee prices at the end of that decade changed this perception. These experiences allowed the actor to recognize the dependence of small producers on the market, as well as the limited institutional presence in the territory.

At the beginning of the new century, the actor joined a regional cooperative as a young technical assistant. Plot visits, record keeping, and meeting organization placed the actor in a space of interaction between technical knowledge and producers' experience. Through this process, the actor understood that assistance did not depend only on transmitting recommendations, but also on building relations of trust. The crisis caused by coffee rust between 2012 and 2014 constituted another turning point, due to production losses among many families and the departure of young people from the community. From that moment onward, the actor oriented their work toward agroforestry systems and, through later training processes, participated in actions related to shade management, soil recovery, and diversification. The arrival of public programs allowed the actor to coordinate a small team and consolidate a position of mediation between institutions, producers, and the community, articulating reports, assemblies, plots, and territorial relations.

The corpus has a lexical density of 0.5576 and a high approximate lexical diversity, indicating discourse that is relatively concentrated in substantive and verbal terms while being organized around a stable set of semantic nuclei. The terms with the greatest weight are coffee, community, technical, and work, followed by producer, plot, family, and institution. This distribution configures a lexical structure in which productive activity appears closely linked to family, community, and institutional relations. Co-occurrences between coffee, plot, parents, and family locate cultivation in the space of early biographical experience, while associations between producer, technical, and institution shift the discourse toward the professional and organizational sphere.

The n-grams reinforce this progression through expressions such as "small coffee plot," "coffee prices," "market dependency," "regional cooperative," "technical assistance," and "community meetings." In the final part of the profile, combinations related to crisis and productive transformation become relevant, including "rust outbreak," "agroforestry systems," "shade management," "soil recovery," and "public programs." The language thus moves from a concrete description of family and plot toward a technical and institutional vocabulary, without abandoning references to community, trust, and territory. The discursive structure maintains a sequential and relational character, in which biographical changes are expressed through connections between actors, productive practices, and spaces of intervention.

The trajectory shows a movement from knowledge of coffee acquired through family belonging and direct participation in work toward a technical position constructed through training, cooperative experience, and interaction with producers. This transition does not imply the replacement of community knowledge by specialized knowledge. On the contrary, the joint recurrence of technical, productive, and community fields indicates that the actor's consolidation depends on their capacity to translate languages, connect scales, and adapt interventions to the conditions of plots and families. Trust emerges as the element that allows technical assistance to become a situated and socially recognized practice.

The productive crisis deepens this movement by transforming the relationship with coffee: from a family and economic resource, it becomes also a territorial problem requiring environmental recovery, diversification, and institutional coordination. In this process, a tension appears between the formal procedures of institutions and the relational dynamics of the community. The actor's final position is configured precisely within that tension,

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

since their work consists of articulating reports and programs with meetings, local experiences, and productive needs. Their consolidation does not derive exclusively from technical expertise, but from a mediating competence that integrates biographical memory, knowledge of plots, and the capacity to negotiate among heterogeneous actors.

Profile Generator Output:

ENT_01

The actor was born in the mid-1980s in a municipal town linked to a coffee-growing region. Their initial experience was shaped by the presence of coffee in the surrounding territory, the observation of markets, the movement of harvests, and the relations between the municipal center and nearby communities. Entry into university expanded this territorial and educational horizon. Around the mid-2000s, the actor joined a student group dedicated to rural development projects, through which they began participating in activities aimed at bringing technical information to communities. This first approach assumed a separation between specialized knowledge and local knowledge, but contact with producers and families showed that productive practices already contained their own forms of observation, comparison, and decision-making.

Later participation in a university extension project constituted a turning point. From that moment onward, the actor's work became less focused on explaining recommendations and more oriented toward listening, organizing information, and accompanying groups. The productive crisis of the early 2010s intensified community meetings and brought forward problems such as plant loss, indebtedness, and migration. After this crisis, the actor continued working with producer groups through participatory diagnosis processes that made it possible to recognize both difficulties and local resources, including springs, shade trees, family labor, seeds, and collective experience. The expansion of public programs consolidated their role in facilitation and systematization, turning the actor into a link between community knowledge and the requirements of organizations and institutions.

The corpus has a lexical density of 0.5604 and a high approximate lexical diversity. Its organization is concentrated around the terms coffee, family, people, work, communities, institutions, and producers, configuring a semantic field in which coffee production is articulated with family, community, and organizational relations. The co-occurrence between coffee and family reinforces the biographical environment in which the activity is initially perceived, while associations between communities, institutions, and information shift the language toward spaces of intervention and mediation. The relations between shade and soil, as well as between producers, soils, and shade, introduce a technical component connected to the direct observation of productive systems.

The n-grams show a sequence that moves from expressions of territorial exposure, such as "municipal town," toward formulations associated with projects, university extension, and technical information. Later, constructions such as "participatory diagnosis processes," "producer groups," "collective experience," "systematize information," and "community processes" appear. This progression reveals a language that moves from technical transmission toward the collective organization of knowledge. The discursive structure maintains a relational character: natural resources, family labor, and productive practices appear as connected components, while institutions are associated with demands for evidence, forms, numbers, and reports.

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

The narrative trajectory and linguistic structure converge in a movement from an initially instructional position toward a function of listening, facilitation, and translation. Knowledge is no longer produced exclusively from the university or institutions; instead, it is constructed through soil comparison, shade observation, risk assessment, and the recovery of producers' accumulated experience. This change also modifies the actor's position, as they no longer present themselves as a transmitter of information but as someone who organizes, articulates, and accompanies collective processes.

In this consolidation, a tension persists between institutional timelines and formats and the duration of community processes. While institutions demand visible results, figures, and rapid evidence, territorial work requires continuity, listening, and recognition of resources that are not always easily quantifiable. The actor's final position is constructed in this intermediate space: their early exposure to a coffee-growing region facilitates understanding of coffee production, their training allows them to engage with technical requirements, and their practice with groups allows them to connect both registers without reducing community knowledge to administrative information.

Source link: <https://chatgpt.com/share/6a5553cf-5e68-83e8-8698-3aaa8fa05203>

PART 04

Representative Output from the Project Timeline Generator GPT

The following material presents a representative output generated by the Project Timeline Generator GPT. This module transforms project-level historical, institutional, socio-environmental, and political references into a structured temporal baseline. The output is used as the temporal framework for subsequent alignment of individual life-history interviews.

Methodological clarification: The project timeline is not generated directly from raw interviews or from an uncured corpus. It requires prior curatorial work by the researcher: delimiting the temporal scope, selecting relevant historical, institutional, socio-environmental, and political references, distinguishing direct input from enriched or standardized events, and deciding which references must be excluded because of incomplete information.

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

Period/Date	National Context (Mexican State)	International Context	Coffee-Growing Territories (Chiapas)	Main Social/Agrarian Programs	Comments on Ideology
1870–1910	Porfirian state; oligarchic modernization; order and progress	European, particularly German, capital; global trade in raw materials	Formation of the plantation system; German investments; debt peonage and labor recruitment; Soconusco and the Sierra Madre as enclave economies	—	Positivism; racialization of Indigenous labor; peasants portrayed as obstacles to progress
1910–1934	Revolution; crisis of the hacienda system; weak state	Withdrawal of European capital; postwar reorganization	Partial breakdown of debt peonage; informal occupation of national lands by plantation laborers; migration from Los Altos	Initial agrarian land grants, limited in coffee-growing areas	Peasants re-emerged as a political problem, but without any structural resolution
1934–1940	Cardenismo; agrarian reform; ejido system	Antifascism; the New Deal	The coffee-sector exception: plantations remained intact, while ejidos played only a marginal role in coffee-growing areas	Agrarian land redistribution: 17.9 million hectares	Peasants conceived as national-popular subjects, although only during a brief political window
1940–1952	Productivist turn under Ávila Camacho and Alemán; end of agrarian reformism	The United States as a development model; technical cooperation; beginning of the Cold War	Consolidation of the plantation–community relationship; tolerated settlements; labor dependence	Mexican Agricultural Program (PAM, 1943); GUANOMEX; crop commissions	Peasants equated with technical backwardness: they had to modernize or be displaced
1950s–1970s	Import-substitution industrialization	Green Revolution; Rockefeller Foundation programs	Expansion of coffee production; dependence on agricultural inputs; no land redistribution	Agricultural extension services; CONASUPO (1965)	Technification presented as the solution; poverty depoliticized

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. Cureus J AI Augment Res 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

1970–1982	Late developmentalist state; state-led mediation	Crisis of the import-substitution industrialization model	Union of Ejido Unions and Solidarity Peasant Groups of Chiapas; La Selva Union of Ejidos; ISMAM (Indigenous People of the Sierra Madre of Motozintla); FIECH (Indigenous Ecological Federation of Chiapas); CIOAC; OCEZ; organizational support from the Diocese of San Cristóbal de Las Casas. Latent agrarian conflicts over national lands and coffee plantations in the Sierra Madre, the Selva region around Ocosingo, and the Frailesca region around Jaltenango	INMECAFÉ; PIDER; COPLAMAR; SAM	Managing poverty without transforming its underlying structures
1982–1988	Debt crisis; structural adjustment	International Monetary Fund and World Bank; Washington Consensus	Falling coffee prices; indebtedness; weakening of INMECAFÉ	Beginning of the dismantling of the agrarian state	Peasants portrayed as a fiscal burden
1989–1994	Consolidated neoliberalism under Salinas	NAFTA; United States hegemony	Coffee crisis; expansion of producer organizations; prelude to the EZLN uprising	PRONASOL (1990); PROCAMPO (1993); PROCEDE (1993)	Targeted assistance; land treated as an asset; peasants redefined as clients
1994–2000	Post-Zapatista period; social counterinsurgency	International cooperation; NGOs	Zapatista autonomy; rejection of the state; massive influx of NGOs	PROGRESA (1997)	Anti-statism and neoliberalism as unlikely allies

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. Cureus J AI Augment Res 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

2000–2006	Neoliberal continuity	“Inclusive” globalization	Professionalization of cooperatives; Fair Trade	Oportunidades (2002)	Poverty framed as a problem of human capital
2007–2012	Evaluative state	Millennium Development Goals agenda	Saturation of the Fair Trade market; exclusion	Continuation of Oportunidades	Technical management of poverty
2013–2018	Late neoliberalism	Financial inclusion	Persistence of monoculture; dependence on development projects	PROAGRO Productivo (2013); Prospera (2014)	Productivity without redistribution; management of the resulting damage

TABLE 3: Example of structured data entered into the timeline generator

Example of structured historical data entered into the timeline generator. The information is organized by period and includes national and international contexts, territorial dynamics, public programs, and dominant ideological frameworks.

Timeline output based upon Table 3:

PROJECT_BASELINE
1870-1910 Porfirian state; oligarchic modernization; order and progress
1870-1910 European capital, mainly German; global trade in raw materials
1870-1910 Formation of the plantation system; German investment; debt peonage and labor recruitment; Soconusco and Sierra Madre as enclaves
1870-1910 Positivism; racialization of Indigenous labor; peasants regarded as obstacles to progress
1910-1934 Revolution; crisis of the hacienda regime; weak state
1910-1934 Withdrawal of European capital; postwar reorganization
1910-1934 Partial breakdown of debt peonage; informal occupation of national lands by laborers; migration from Los Altos
1910-1934 First agrarian land grants, limited in coffee-growing territories
1910-1934 Peasants as a political problem without structural resolution
1934-1940 Cardenismo; agrarian reform; ejido
1934-1940 Antifascism; New Deal
1934-1940 Coffee-growing exception: plantations remained in place and ejidos had a marginal presence in coffee-growing areas
1934-1940 Agrarian redistribution of 17.9 million hectares
1934-1940 Peasants as national-popular subjects
1940-1952 Productivist shift under the Ávila Camacho and Alemán administrations; end of agrarian reform
1940-1952 United States as a model; technical cooperation; beginning of the Cold War
1940-1952 Consolidation of the plantation-community relationship; tolerated settlements; labor dependency

How to cite this article:

1940-1952 | PAM
1940-1952 | GUANOMEX
1940-1952 | Crop commissions
1940-1952 | Peasants regarded as technically backward; modernize or displace
1950-1979 | Import-substitution industrialization
1950-1979 | Green Revolution; Rockefeller foundations
1950-1979 | Expansion of coffee cultivation; dependence on agricultural inputs; absence of land redistribution
1950-1979 | Agricultural extension
1950-1979 | CONASUPO
1950-1979 | Technification as the solution; depoliticization of poverty
1970-1982 | Late developmentalist state; state mediation
1970-1982 | Crisis of the import-substitution industrialization model
1970-1982 | Union of Ejido Unions and Solidarity Peasant Groups of Chiapas
1970-1982 | La Selva Union of Ejidos
1970-1982 | ISMAM
1970-1982 | FIECH
1970-1982 | CIOAC
1970-1982 | OCEZ
1970-1982 | Organizational support from the Diocese of San Cristóbal de Las Casas
1970-1982 | Latent agrarian conflicts over national lands and coffee plantations in Sierra Madre, the Ocosingo forest region, and the Jaltenango area of Frailesca
1970-1982 | INMECAFÉ
1970-1982 | PIDER
1970-1982 | COPLAMAR
1970-1982 | SAM
1970-1982 | Management of poverty without changing the underlying structure
1982-1988 | Debt crisis; structural adjustment
1982-1988 | IMF; World Bank; Washington Consensus
1982-1988 | Falling prices; indebtedness; weakening of INMECAFÉ
1982-1988 | Beginning of the dismantling of the agrarian state
1982-1988 | Peasants regarded as a fiscal burden
1989-1994 | Full neoliberalism under the Salinas administration
1989-1994 | NAFTA; United States hegemony
1989-1994 | Coffee crisis; organizational expansion; prelude to the EZLN
1989-1994 | PRONASOL
1989-1994 | PROCAMPO
1989-1994 | PROCEDE
1989-1994 | Targeting; land as an asset; peasants regarded as clients
1994-2000 | Post-Zapatista period; social counterinsurgency
1994-2000 | International cooperation; nongovernmental organizations

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

1994-2000 | Zapatista autonomy; rejection of the state; large-scale arrival of nongovernmental organizations

1994-2000 | PROGRESA

1994-2000 | Anti-statism and neoliberalism as unlikely allies

2000-2006 | Neoliberal continuity

2000-2006 | Inclusive globalization

2000-2006 | Professionalization of cooperatives; fair trade

2000-2006 | Oportunidades

2000-2006 | Poverty regarded as a human-capital problem

2007-2012 | Evaluative state

2007-2012 | Millennium Development Goals agenda

2007-2012 | Saturation of fair trade; exclusion

2007-2012 | Continuation of Oportunidades

2007-2012 | Technical management of poverty

2013-2018 | Late neoliberalism

2013-2018 | Financial inclusion

2013-2018 | Persistence of monoculture; dependence on projects

2013-2018 | PROAGRO Productivo

2013-2018 | Prospera

2013-2018 | Productivity without redistribution; management of harm

EVENT_ORIGIN

Porfirian state; oligarchic modernization; order and progress | direct_input

European capital, mainly German; global trade in raw materials | direct_input

Formation of the plantation system; German investment; debt peonage and labor recruitment; Soconusco and Sierra Madre as enclaves | direct_input

Positivism; racialization of Indigenous labor; peasants regarded as obstacles to progress | direct_input

Revolution; crisis of the hacienda regime; weak state | direct_input

Withdrawal of European capital; postwar reorganization | direct_input

Partial breakdown of debt peonage; informal occupation of national lands by laborers; migration from Los Altos | direct_input

First agrarian land grants, limited in coffee-growing territories | direct_input

Peasants as a political problem without structural resolution | direct_input

Cardenismo; agrarian reform; ejido | direct_input

Antifascism; New Deal | direct_input

Coffee-growing exception: plantations remained in place and ejidos had a marginal presence in coffee-growing areas | direct_input

Agrarian redistribution of 17.9 million hectares | direct_input

Peasants as national-popular subjects | direct_input

Productivist shift under the Ávila Camacho and Alemán administrations; end of agrarian reform | direct_input

United States as a model; technical cooperation; beginning of the Cold War | direct_input

Consolidation of the plantation-community relationship; tolerated settlements; labor dependency | direct_input

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

PAM | direct_input
GUANOMEX | direct_input
Crop commissions | direct_input
Peasants regarded as technically backward; modernize or displace | direct_input
Import-substitution industrialization | direct_input
Green Revolution; Rockefeller foundations | direct_input
Expansion of coffee cultivation; dependence on agricultural inputs; absence of land redistribution | direct_input
Agricultural extension | direct_input
CONASUPO | direct_input
Technification as the solution; depoliticization of poverty | direct_input
Late developmentalist state; state mediation | direct_input
Crisis of the import-substitution industrialization model | direct_input
Union of Ejido Unions and Solidarity Peasant Groups of Chiapas | direct_input
La Selva Union of Ejidos | direct_input
ISMAM | direct_input
FIECH | direct_input
CIOAC | direct_input
OCEZ | direct_input
Organizational support from the Diocese of San Cristóbal de Las Casas | direct_input
Latent agrarian conflicts over national lands and coffee plantations in Sierra Madre, the Ocosingo forest region, and the Jaltenango area of Frailesca | direct_input
INMECAFÉ | direct_input
PIDER | direct_input
COPLAMAR | direct_input
SAM | direct_input
Management of poverty without changing the underlying structure | direct_input
Debt crisis; structural adjustment | direct_input
IMF; World Bank; Washington Consensus | direct_input
Falling prices; indebtedness; weakening of INMECAFÉ | direct_input
Beginning of the dismantling of the agrarian state | direct_input
Peasants regarded as a fiscal burden | direct_input
Full neoliberalism under the Salinas administration | direct_input
NAFTA; United States hegemony | direct_input
Coffee crisis; organizational expansion; prelude to the EZLN | direct_input
PRONASOL | direct_input
PROCAMPO | direct_input
PROCEDE | direct_input
Targeting; land as an asset; peasants regarded as clients | direct_input
Post-Zapatista period; social counterinsurgency | direct_input
International cooperation; nongovernmental organizations | direct_input

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

Zapatista autonomy; rejection of the state; large-scale arrival of nongovernmental organizations	direct_input
PROGRESA	direct_input
Anti-statism and neoliberalism as unlikely allies	direct_input
Neoliberal continuity	direct_input
Inclusive globalization	direct_input
Professionalization of cooperatives; fair trade	direct_input
Oportunidades	direct_input
Poverty regarded as a human-capital problem	direct_input
Evaluative state	direct_input
Millennium Development Goals agenda	direct_input
Saturation of fair trade; exclusion	direct_input
Continuation of Oportunidades	direct_input
Technical management of poverty	direct_input
Late neoliberalism	direct_input
Financial inclusion	direct_input
Persistence of monoculture; dependence on projects	direct_input
PROAGRO Productivo	direct_input
Prospera	direct_input
Productivity without redistribution; management of harm	direct_input
ENRICHED_EVENTS	
1950s-1970s → 1950-1979	date_normalization
EXCLUDED_EVENTS	
Absence of a social or agrarian program for 1870-1910	incomplete_information
TEMPORAL_COVERAGE	
1870 2018	
EVENT_DENSITY	
1870-1910	4
1910-1934	5
1934-1940	5
1940-1952	7
1950-1979	6
1970-1982	15
1982-1988	5
1989-1994	7
1994-2000	5
2000-2006	5
2007-2012	5
2013-2018	6

How to cite this article:

Source link: <https://chatgpt.com/share/6a555695-3b3c-83e8-83bd-5c296272b040>

PART 05

Representative Output from the Project Timeline Generator GPT

Representative Output from the Temporal Alignment GPT

The following material presents representative outputs generated by the Temporal Alignment GPT for two synthetic life-history interviews. In this stage, the module receives two inputs: (1) the general project temporal baseline and (2) the cleaned transcript of each life history. The module does not reinterpret the historical baseline and does not modify explicit biographical dates. Its function is to locate biographical events in time, classify the type and precision of temporal anchoring, relate each event to the project baseline, and document uncertainty.

These examples are synthetic and are included only to demonstrate how the temporal alignment stage operates within the Orchestrated Distant Reading workflow.

Temporal Alignment Output:

TEMPORAL_ALIGNMENT
Birth in a municipal town where coffee was part of the surrounding landscape 1985 exact explicit high
Childhood familiarity with coffee through neighbors, markets, trucks, and seasonal movement from communities late 1980s-early 2000s range derived_by_age medium
Entry into university and initial involvement in rural development work around 2006 range explicit high
Participation in a student group working on rural development projects around 2006-2010 range derived_by_sequence medium
Initial belief that communities primarily needed external technical information around 2006-2010 range derived_by_sequence medium
Recognition that communities already possessed knowledge not acknowledged by institutions around 2006-2010 range derived_by_sequence medium
Participation in a university extension project with coffee producers 2010 exact explicit high
Change from an instructional approach toward recognition of producers' experimentation and decision-making 2010 onward range derived_by_sequence medium
Development of a listening, information-organizing, and comparative support role 2010-2012 range derived_by_sequence medium
Experience of production decline and plant loss during the coffee crisis 2012-2014 range explicit high
Participation in intensified meetings about varieties, shade, pruning, soil, and the continuity of coffee production 2012-2014 range derived_by_sequence high
Recognition that the coffee crisis affected family decisions, migration, debt, and institutional trust 2012-2014 range derived_by_sequence high
Continuation of work with producer groups through participatory diagnosis processes after 2015 relative explicit high
Mapping of community problems and resources, including springs, shade trees, family labor, local seed, and collective experience after 2015-2019 range derived_by_sequence medium
Support for organizations systematizing information after the expansion of public programs 2019 exact explicit high

How to cite this article:

Collaboration with teams seeking to understand conditions in coffee-producing communities | 2019 onward | range | derived_by_sequence | medium

Consolidation of a facilitation role connecting community knowledge with institutional requirements | 2019-present | range | derived_by_sequence | medium

Ongoing effort to prevent community experience from being reduced to forms, numbers, or reports | 2019-present | range | derived_by_sequence | medium

Recognition of the tension between rapid institutional evidence and the slower temporality of community processes | present | relative | derived_by_sequence | medium

BASELINE_MATCH

Birth in a municipal town where coffee was part of the surrounding landscape | Debt crisis; structural adjustment | contemporary

Childhood familiarity with coffee through neighbors, markets, trucks, and seasonal movement from communities |

Coffee crisis; organizational growth; prelude to the EZLN | contemporary

Entry into university and initial involvement in rural development work | Professionalization of cooperatives; fair trade | contemporary

Participation in a student group working on rural development projects | Professionalization of cooperatives; fair trade | contemporary

Initial belief that communities primarily needed external technical information | Poverty considered a human-capital problem | contemporary

Recognition that communities already possessed knowledge not acknowledged by institutions | Professionalization of cooperatives; fair trade | contemporary

Participation in a university extension project with coffee producers | Technical management of poverty | contemporary

Change from an instructional approach toward recognition of producers' experimentation and decision-making | Technical management of poverty | contemporary

Development of a listening, information-organizing, and comparative support role | Evaluative state | contemporary

Experience of production decline and plant loss during the coffee crisis | Persistence of monoculture; dependence on projects | contemporary

Participation in intensified meetings about varieties, shade, pruning, soil, and the continuity of coffee production | Productivity without redistribution; management of damage | contemporary

Recognition that the coffee crisis affected family decisions, migration, debt, and institutional trust | Productivity without redistribution; management of damage | contemporary

Continuation of work with producer groups through participatory diagnosis processes | PROAGRO Productivo; Prospera | contemporary

Mapping of community problems and resources, including springs, shade trees, family labor, local seed, and collective experience | Persistence of monoculture; dependence on projects | contemporary

Support for organizations systematizing information after the expansion of public programs | Productivity without redistribution; management of damage | later

Collaboration with teams seeking to understand conditions in coffee-producing communities | Productivity

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

without redistribution; management of damage | later

Consolidation of a facilitation role connecting community knowledge with institutional requirements |

Productivity without redistribution; management of damage | later

Ongoing effort to prevent community experience from being reduced to forms, numbers, or reports | Productivity without redistribution; management of damage | later

Recognition of the tension between rapid institutional evidence and the slower temporality of community processes | Productivity without redistribution; management of damage | later

TEMPORAL_SEQUENCE

Birth in a municipal town where coffee was part of the surrounding landscape → Childhood familiarity with coffee through neighbors, markets, trucks, and seasonal movement from communities

Childhood familiarity with coffee through neighbors, markets, trucks, and seasonal movement from communities → Entry into university and initial involvement in rural development work

Entry into university and initial involvement in rural development work → Participation in a student group working on rural development projects

Participation in a student group working on rural development projects → Initial belief that communities primarily needed external technical information

Initial belief that communities primarily needed external technical information → Recognition that communities already possessed knowledge not acknowledged by institutions

Recognition that communities already possessed knowledge not acknowledged by institutions → Participation in a university extension project with coffee producers

Participation in a university extension project with coffee producers → Change from an instructional approach toward recognition of producers' experimentation and decision-making

Change from an instructional approach toward recognition of producers' experimentation and decision-making → Development of a listening, information-organizing, and comparative support role

Development of a listening, information-organizing, and comparative support role → Experience of production decline and plant loss during the coffee crisis

Experience of production decline and plant loss during the coffee crisis → Participation in intensified meetings about varieties, shade, pruning, soil, and the continuity of coffee production

Participation in intensified meetings about varieties, shade, pruning, soil, and the continuity of coffee production → Recognition that the coffee crisis affected family decisions, migration, debt, and institutional trust

Recognition that the coffee crisis affected family decisions, migration, debt, and institutional trust → Continuation of work with producer groups through participatory diagnosis processes

Continuation of work with producer groups through participatory diagnosis processes → Mapping of community problems and resources, including springs, shade trees, family labor, local seed, and collective experience

Mapping of community problems and resources, including springs, shade trees, family labor, local seed, and collective experience → Support for organizations systematizing information after the expansion of public programs

Support for organizations systematizing information after the expansion of public programs → Collaboration with teams seeking to understand conditions in coffee-producing communities

Collaboration with teams seeking to understand conditions in coffee-producing communities → Consolidation of

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

a facilitation role connecting community knowledge with institutional requirements

Consolidation of a facilitation role connecting community knowledge with institutional requirements → Ongoing effort to prevent community experience from being reduced to forms, numbers, or reports

Ongoing effort to prevent community experience from being reduced to forms, numbers, or reports →

Recognition of the tension between rapid institutional evidence and the slower temporality of community processes

TEMPORAL_LAG

Birth in a municipal town where coffee was part of the surrounding landscape | Debt crisis; structural adjustment | contemporary

Childhood familiarity with coffee through neighbors, markets, trucks, and seasonal movement from communities |

Coffee crisis; organizational growth; prelude to the EZLN | contemporary

Entry into university and initial involvement in rural development work | Professionalization of cooperatives; fair trade | contemporary

Participation in a student group working on rural development projects | Professionalization of cooperatives; fair trade | contemporary

Initial belief that communities primarily needed external technical information | Poverty considered a human-capital problem | contemporary

Recognition that communities already possessed knowledge not acknowledged by institutions | Professionalization of cooperatives; fair trade | contemporary

Participation in a university extension project with coffee producers | Technical management of poverty | contemporary

Change from an instructional approach toward recognition of producers' experimentation and decision-making | Technical management of poverty | contemporary

Development of a listening, information-organizing, and comparative support role | Evaluative state | contemporary

Experience of production decline and plant loss during the coffee crisis | Persistence of monoculture; dependence on projects | contemporary

Participation in intensified meetings about varieties, shade, pruning, soil, and the continuity of coffee production |

Productivity without redistribution; management of damage | contemporary

Recognition that the coffee crisis affected family decisions, migration, debt, and institutional trust | Productivity without redistribution; management of damage | contemporary

Continuation of work with producer groups through participatory diagnosis processes | PROAGRO Productivo; Prospera | contemporary

Mapping of community problems and resources, including springs, shade trees, family labor, local seed, and collective experience | Persistence of monoculture; dependence on projects | contemporary

Support for organizations systematizing information after the expansion of public programs | Productivity without redistribution; management of damage | later

Collaboration with teams seeking to understand conditions in coffee-producing communities | Productivity without redistribution; management of damage | later

Consolidation of a facilitation role connecting community knowledge with institutional requirements |

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

Productivity without redistribution; management of damage | later

Ongoing effort to prevent community experience from being reduced to forms, numbers, or reports | Productivity without redistribution; management of damage | later

Recognition of the tension between rapid institutional evidence and the slower temporality of community processes | Productivity without redistribution; management of damage | later

TEMPORAL_UNCERTAINTY

Childhood familiarity with coffee through neighbors, markets, trucks, and seasonal movement from communities | missing_explicit_date | The experience is located in childhood, but no exact beginning or ending year is given.

Entry into university and initial involvement in rural development work | relative_reference | "Around 2006" identifies an approximate year rather than an exact date.

Participation in a student group working on rural development projects | missing_explicit_date | Participation began around 2006, but its duration is not specified.

Initial belief that communities primarily needed external technical information | missing_explicit_date | The belief belongs to the initial stage of university involvement, but no precise period is stated.

Recognition that communities already possessed knowledge not acknowledged by institutions | relative_reference | The change occurred after the initial technical approach but before or around the 2010 extension project.

Change from an instructional approach toward recognition of producers' experimentation and decision-making | missing_explicit_date | The transformation followed the 2010 project, but its duration and exact point of consolidation are not specified.

Development of a listening, information-organizing, and comparative support role | missing_explicit_date | The role emerged after the 2010 project without a precise starting year.

Continuation of work with producer groups through participatory diagnosis processes | relative_reference | "After 2015" establishes a lower temporal boundary but no exact starting date or endpoint.

Mapping of community problems and resources, including springs, shade trees, family labor, local seed, and collective experience | missing_explicit_date | The activity occurred after 2015 and before or through 2019, but its precise duration is not stated.

Collaboration with teams seeking to understand conditions in coffee-producing communities | missing_explicit_date | The collaboration is linked to 2019 but has no explicit endpoint.

Consolidation of a facilitation role connecting community knowledge with institutional requirements | missing_explicit_date | The narrative presents facilitation as a continuing role without a precise date of consolidation.

Ongoing effort to prevent community experience from being reduced to forms, numbers, or reports | missing_explicit_date | The activity is described as part of the narrator's current practice but lacks an explicit starting date.

Recognition of the tension between rapid institutional evidence and the slower temporality of community processes | relative_reference | The observation belongs to the narrator's present perspective, but no calendar date is provided.

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

Source link: <https://chatgpt.com/share/6a5558d1-f6b8-83e8-a921-373bd0b213d6>

PART 06

Representative Output from the Cross-Case Analysis GPT

The following material presents a representative output generated by the Cross-Case Analysis GPT after processing two synthetic life-history profiles, their structured analytical outputs, and their temporal alignments. The output illustrates how the module identifies recurrent patterns, relevant contrasts, key articulations, temporal dynamics, common semantic nuclei, territorial variations, structural indicators, partial findings, and limits of the comparative block.

COMPARATIVE_ANALYSIS_BLOCK
THEME
Transformation of technical and mediation functions in coffee-growing communities in response to production crises, organizational processes, and institutional demands.
QUESTION
How are mediation trajectories among producers, communities, and institutions comparatively configured, and what changes do they undergo during coffee crises and the expansion of public programs?
CASES_ANALYZED
ENT_01
ENT_02
RECURRING_PATTERNS
FORMAT:
pattern description evidence
biographical_connection_to_coffee In both cases, coffee is part of the family, community, and territorial environment before the individuals assume technical functions. ENT_01: coffee-growing family, landscape, markets, and seasonal movements; ENT_02: family plot, harvesting, kinship, and community belonging.
transition_from_technical_approach_to_mediation Both trajectories move from functions centered on transmitting or translating technical information toward listening, accompaniment, and articulation practices. ENT_01 moves from delivering technical information to organizing information and facilitating comparisons; ENT_02 moves from translating technical language to understanding assistance as a relationship of trust.
recognition_of_producer_knowledge In both cases, producers are recognized as actors who experiment, compare resources, and make decisions about their plots. ENT_01 records comparisons of soils, observation of shade, and assessment of risks; ENT_02 links technical assistance to visits, meetings, and changes developed through trust.
2012_2014_crisis_as_turning_point The production crisis concentrates material losses, intensified meetings, and a reorientation of technical work. ENT_01 records production decline, plant loss, migration, debt, and discussions about varieties, pruning, shade, and soil; ENT_02 records production losses, youth outmigration, and subsequent interest in agroforestry systems.
continuity_of_post_crisis_collective_work After the crisis, mediation functions continue through diagnosis, training, accompaniment, and the organization of resources and information. ENT_01 maintains participatory

How to cite this article:

diagnoses and resource mapping after 2015; ENT_02 participates after 2016 in processes involving shade, soil recovery, and diversification.

expansion_of_functions_in_2019 | The expansion or arrival of public programs broadens responsibilities for coordination among producers, organizations, and institutions. | ENT_01 supports the systematization of information for organizations and teams; ENT_02 assumes responsibility for coordinating an accompaniment team.

tension_between_community_processes_and_institutional_requirements | Mediation connects long-term territorial processes with institutional demands for reports, forms, and evidence. | ENT_01 explicitly contrasts demands for rapid evidence with community timelines; ENT_02 mediates among reports, assemblies, institutions, and work on plots.

RELEVANT_CONTRASTS

FORMAT:

contrast | description | cases

origin_of_mediation_legitimacy | ENT_01 builds a mediation position through university education, extension work, and rural projects; ENT_02 builds it through family experience in coffee production and work in a regional cooperative. | ENT_01, ENT_02

moment_of_professional_reorientation | In ENT_01, the shift toward listening is concentrated around the 2010 university project; in ENT_02, assistance based on trust develops gradually from cooperative work beginning around 2001. | ENT_01, ENT_02

initial_mode_of_intervention | ENT_01 begins with information transmission and an instructional position; ENT_02 begins with the translation of technical language, plot visits, and meeting organization. | ENT_01, ENT_02

highlighted_social_effect_of_the_crisis | ENT_01 emphasizes family decisions, migration, debt, and institutional trust; ENT_02 emphasizes production loss and youth departure from the community. | ENT_01, ENT_02

post_crisis_technical_response | ENT_01 focuses on participatory diagnosis and mapping problems and resources; ENT_02 focuses on agroforestry systems, shade, soil recovery, and diversification. | ENT_01, ENT_02

institutional_function_since_2019 | ENT_01 concentrates on systematizing information and connecting community knowledge with institutional requests; ENT_02 assumes team coordination and mediation among technical, organizational, and territorial spaces. | ENT_01, ENT_02

territorial_configuration | ENT_01 represents territory through communities, resources, and information flows; ENT_02 represents it through plots, relationships, memory, and institutions. | ENT_01, ENT_02

KEY_ARTICULATIONS

FORMAT:

articulation | actors | events | territory | temporality

technical_and_community_learning | producers, communities, technicians, university, and cooperative | rural projects, university extension, plot visits, meetings, and technical translation | communities, plots, and the coffee-growing region | 2001-2012

production_crisis_and_reorganization | families, producers, young people, and institutions | production decline, plant loss, coffee rust outbreak, and intensified meetings | coffee-growing communities and plots | 2012-2014

post_crisis_socioenvironmental_management | producers, producer groups, and technical teams | participatory

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

diagnosis, shade management, soil recovery, diversification, and resource mapping | plots, soils, springs, shade trees, and communities | after 2015

programmatic_mediation | public programs, organizations, institutions, teams, and producers | information systematization, team coordination, and accompaniment | communities, assemblies, institutional spaces, and plots | since 2019

translation_between_knowledge_and_requirements | communities, producers, technicians, and institutions | listening, information organization, report preparation, and knowledge connection | coffee-growing territory and institutional spaces | continuity from the initial technical experience

TEMPORAL_DYNAMICS

FORMAT:

temporal_pattern | description | type

convergence_during_the_2012_2014_crisis | Both cases identify the 2012-2014 production crisis as a period of material loss, organizational intensification, and redefinition of technical work. | concentration

gradual_consolidation_of_mediation | Mediation is configured cumulatively through projects, meetings, visits, listening, and continued work with producers. | continuity

difference_in_entry_into_technical_work | ENT_02 begins cooperative work around 2001, while ENT_01 enters rural projects around 2006 and concentrates its reorientation in 2010. | time_lag

post_crisis_reorientation | Management, training, and diagnosis responses appear after the crisis: after 2015 in ENT_01 and after 2016 in ENT_02. | time_lag

institutional_expansion_in_2019 | In both cases, 2019 concentrates new functions related to public programs, teams, and the production of institutional information. | concentration

break_with_the_instructional_position | The cases show a movement away from functions centered exclusively on instructing, explaining, or transmitting technical solutions. | discontinuity

persistence_of_links_with_producers | Despite changes in position, institution, and program, work with producers, groups, and communities continues. | continuity

mismatch_between_institutional_and_community_timelines | ENT_01 explicitly identifies the difference between rapid demands for evidence and prolonged community processes; ENT_02 continuously connects reports, assemblies, and plot-based work. | time_lag

COMMON_SEMANTIC_CORES

FORMAT:

term | relative_weight | case_distribution

coffee | high | ENT_01: 0.0321; ENT_02: 0.0235

work | high | ENT_01: 0.0229; ENT_02: 0.0235

community_communities | high | ENT_01: 0.0183; ENT_02: 0.0235

producers_producer | high | ENT_01: 0.0183; ENT_02: 0.0188

family | medium_high | ENT_01: 0.0229; ENT_02: 0.0141

institutions_institution | medium_high | ENT_01: 0.0183; ENT_02: 0.0141

technical | concentrated_in_ENT_02 | ENT_02: 0.0235; present in ENT_01 through "technical information"

people | concentrated_in_ENT_01 | ENT_01: 0.0229; ENT_02: 0.0094

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

shade | shared_with_greater_weight_in_ENT_01 | ENT_01: 0.0138; ENT_02: present through “shade management” information | concentrated_in_ENT_01 | ENT_01: 0.0138; ENT_02: present through notes and reports
trust | concentrated_in_ENT_02 | ENT_02: 0.0094; thematically present in ENT_01 through institutional trust
soil_soils | shared | ENT_01: comparison and mapping; ENT_02: recovery and the relationship between shade and soil

TERRITORIAL_VARIATIONS

FORMAT:

territory | pattern | cases

municipal_town_and_surrounding_communities | The connection with coffee is configured through landscape, markets, transportation, and seasonal movements between the municipal town and communities. | ENT_01

small_highland_community | The territorial connection is configured through residence, the family plot, harvesting, kinship, and direct community work. | ENT_02

coffee_plots | They function as spaces for observation, experimentation, accompaniment, and trust-building. | ENT_01, ENT_02

communities | They constitute the primary spaces for meetings, learning, diagnosis, and articulation with institutions. | ENT_01, ENT_02

region | It appears as the scale through which cooperatives, programs, and technical work circulate, although with limited spatial specificity. | ENT_01, ENT_02

territory_as_relational_network | It is articulated through memory, relationships, community, plots, and institutions. | ENT_02

territory_as_resource_inventory | It is expressed through the mapping of springs, shade trees, family labor, local seed, and collective experience. | ENT_01

STRUCTURAL_INDICATORS

FORMAT:

indicator | tendency | evidence

lexical_density | stable_and_similar | ENT_01: 0.5604; ENT_02: 0.5576

approximate_lexical_diversity | high_in_both_cases | ENT_01: 0.7844; ENT_02: 0.7981

semantic_density | slightly_higher_in_ENT_02 | ENT_01: 0.1972; ENT_02: 0.2066

number_of_actors | higher_in_ENT_01 | ENT_01: 26; ENT_02: 23

number_of_events | similar | ENT_01: 14; ENT_02: 13

centrality_of_coffee | higher_in_ENT_01 | ENT_01: 0.0321; ENT_02: 0.0235

centrality_of_institutional_mediation | present_in_both_with_different_functions | ENT_01: information, evidence, and institutional requests; ENT_02: technical translation, trust, coordination, and reports

PARTIAL_FINDINGS

FORMAT:

finding

Mediation trajectories are not configured as fixed positions, but as cumulative functions developed through educational, organizational, and territorial experiences.

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

The 2012-2014 crisis constitutes the main shared temporal turning point in the redefinition of technical practices and the intensification of collective work.

The central transformation consists of a shift from knowledge transmission toward listening, comparison, accompaniment, and translation between positions.

The expansion of public programs in 2019 does not replace previous community functions; instead, it adds team coordination, information systematization, and the production of institutional evidence.

Differences in professional origin produce distinct forms of mediation: one more closely linked to extension, diagnosis, and information, and the other to cooperative work, plots, trust, and territorial coordination.

Plots and communities remain central spaces for knowledge production, while programs expand mediation into reports, organizations, and institutional requirements.

BLOCK_LIMITATIONS

FORMAT:

limitation

The available material contains only two distinct cases, ENT_01 and ENT_02, and therefore does not allow recurring patterns to be established across a set of three to five trajectories.

The cases share general temporal references but do not provide the same level of detail for each event, institution, or program.

Territorial references are broad and do not permit precise comparisons among specific communities, municipalities, or regions.

The data support the identification of associations, sequences, and contrasts, but not strong causal attributions concerning professional or institutional changes.

Source link: <https://chatgpt.com/share/6a580c40-2118-83e8-9096-7804aac6eb1d>

PART 07

Representative Final Output from the Cross-Case Synthesis GPT

The following material presents a representative final output generated after the cross-case analysis of two synthetic life-history interviews. It illustrates how the workflow can transform structured outputs, temporal alignments, narrative profiles, and comparative blocks into a controlled analytical synthesis. The cases are synthetic and are included only to demonstrate the progression of the Orchestrated Distant Reading workflow.

RESULTS

The trajectories analyzed show a convergent transformation of technical and mediation functions in coffee-growing communities. In both cases, coffee constitutes an early component of the biographical, family, and territorial environment, although the initial connection takes different forms. In ENT_01, the relationship is configured from a municipal town through the landscape, markets, transportation, and exchanges with surrounding communities. In ENT_02, it is constructed within a small highland community through the family plot, harvest work, kinship, and direct community participation. These differences in territorial origin do not prevent both trajectories from later converging in functions of accompaniment, facilitation, and institutional articulation.

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

The pathways into technical work are also differentiated. ENT_01 enters through the university, a student group, and university extension activities that began around 2006. ENT_02 begins their professional trajectory through a regional cooperative, after a family and educational experience linked to coffee production, and starts cooperative work around 2001. This temporal lag shows that mediation functions do not depend on a single pathway of formation. They can develop both from university spaces and from cooperative structures and prior productive experiences.

In both cases, technical work undergoes a shift from the transmission or translation of knowledge toward practices centered on listening, trust, and information organization. In ENT_01, this reorientation is concentrated around the 2010 university project, when the function ceased to consist mainly of bringing technical information and began to include facilitating comparisons and organizing the knowledge expressed by producers. In ENT_02, the change develops gradually during cooperative work, through the translation of technical language and the understanding of assistance as a relationship constructed through trust and direct work on plots.

This change in professional position is accompanied by a growing recognition of producer knowledge. Producers do not appear only as recipients of instructions, but as actors who observe, experiment, compare, and make decisions. In ENT_01, this recognition is expressed through practices related to soils, shade, seeds, risks, and community resources. In ENT_02, it is expressed through the relationship between changes made in the plots, the trust built, and technical accompaniment. Coffee plots therefore function as shared spaces of observation, experimentation, learning, and production of trust.

The productive crisis of 2012-2014 constitutes the main shared turning point. Both cases register a fall in production, plant loss, and intensification of collective work. ENT_01 highlights family decisions related to migration, debt, and trust in institutions, together with meetings on varieties, shade, pruning, and soil management. ENT_02 emphasizes productive losses and the departure of young people from the community, as well as later interest in agroforestry systems. The crisis concentrates material losses and social changes, but it also expands meetings, comparisons between practices, and the reorganization of technical and community work.

Post-crisis responses combine productive, socio-environmental, and organizational dimensions. After 2015, ENT_01 participates in community diagnoses and resource mapping involving springs, shade trees, local seeds, family labor, and collective experience. After 2016, ENT_02 becomes involved in training activities on agroforestry systems, shade management, soil recovery, and diversification. In both cases, work with producer groups continues, but the orientations differ: ENT_01 privileges participatory diagnosis and the organization of territorial information, while ENT_02 focuses on the recovery and diversification of plots.

Socio-environmental resources occupy a transversal position in the trajectories. Soil, shade, seeds, springs, trees, and plots are objects of management, but also supports for learning, memory, and collective organization. In ENT_01, territory is presented mainly as a set of communities, resources, and information circuits that can be identified and systematized. In ENT_02, it appears as a network of plots, relationships, community memories, and institutions. This variation does not modify the centrality of communities and plots as spaces of meeting, accompaniment, and articulation.

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

The expansion of public programs in 2019 broadens and formalizes mediation functions that had been developed previously. ENT_01 begins to support the systematization of information for organizations and institutional teams, connecting community knowledge with external requests and formats. ENT_02 assumes the coordination of a team accompanying producers and articulates institutions, assemblies, reports, and work on plots. In neither case does mediation begin with public programs. Rather, these programs reuse capacities previously formed in the university, the cooperative, plot visits, meetings, and community work.

The structural indicators support the proximity between the two trajectories. Lexical density is stable and similar, with values of 0.5604 in ENT_01 and 0.5576 in ENT_02, while approximate lexical diversity is high in both cases: 0.7844 and 0.7981, respectively. Semantic density is slightly higher in ENT_02, with 0.2066 compared to 0.1972 in ENT_01. The number of actors is higher in ENT_01-26 compared to 23-while the number of events is similar-14 and 13. These proximities show a convergent complexity in the articulation of families, producers, institutions, productive crises, resources, and community spaces.

The common semantic nuclei reinforce the centrality of coffee, work, communities, producers, families, and institutions. The term associated with coffee has greater relative frequency in ENT_01, with 0.0321 compared to 0.0235 in ENT_02. The technical dimension has greater weight in ENT_02, where it reaches 0.0235 and is linked to assistance, language, and work. Information is concentrated in ENT_01, while trust has greater presence in ENT_02. Soil and shade appear in both cases, although they are linked to different practices of comparison, mapping, recovery, and management.

Taken together, the trajectories show a growing expansion of professional functions. The trajectory begins with learning, assistance, or technical transmission; continues with listening, translation, and accompaniment; and later expands toward team coordination, information systematization, and articulation among producers, communities, and institutions. Despite changes in role, program, and space of intervention, the relationship with producers and community groups remains constant.

EMPIRICAL DISCUSSION

The comparison shows that mediation trajectories are configured through the accumulation of experiences rather than through an initially defined function. University projects, cooperatives, meetings, plot visits, and work with producers form a sequence in which technical tasks progressively expand. Mediation is consolidated when the transmission of knowledge ceases to be the exclusive activity and becomes combined with listening, translation, trust, and information organization.

Differences in territorial origin and pathways into professional work do not produce opposing trajectories. ENT_01 begins from a territorial relationship constructed between the municipal town and the communities, and later enters through university spaces. ENT_02 begins from family productive experience and enters through a cooperative. Despite these differences, both cases come to occupy intermediate positions between producers, communities, and institutions. This convergence is observed in the continuity of accompaniment, the recognition of producer knowledge, and the expansion of institutional responsibilities.

The 2012-2014 crisis concentrates the main shared change. Productive losses are related to migration, debt, the departure of young people, changes in institutional trust, and the reorganization of meetings and practices. However, each case highlights different dimensions of the same period. ENT_01 emphasizes family decisions and

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

relations with institutions, while ENT_02 emphasizes production loss and the departure of young people. The crisis functions as a temporal concentration of productive, family, community, and organizational processes, although the available information does not allow causal relationships to be established among them.

Post-crisis responses show that technical work shifts toward forms of management that incorporate social, environmental, and territorial resources. Participatory diagnosis and resource mapping in ENT_01 organize information on collective and community elements. Agroforestry systems, shade management, soil recovery, and diversification in ENT_02 focus on plot recovery. Both orientations maintain the link with producers, but they differ in the scale and type of practice privileged.

The institutional expansion of 2019 introduces new responsibilities without replacing previously constructed capacities. Tasks such as systematization, report writing, team coordination, and programmatic articulation depend on prior experiences of listening, translation, and community work. Public programs formalize mediation and expand its spaces of exercise toward offices, organizations, and institutional teams, while maintaining a presence in communities, assemblies, and plots.

This expansion generates a tension between institutional requirements and community processes. ENT_01 explicitly contrasts the demand for rapid evidence with the slower temporality of community dynamics. ENT_02 presents continuous mediation among reports, assemblies, and work on plots. In both cases, the production of institutional information requires translating territorial experiences into external formats, although the rhythms of programs, reports, and communities are not fully synchronized.

Within the analyzed cases, the research question is answered through the identification of trajectories that move from initial technical functions toward positions of articulation. These trajectories are configured through biographical links with coffee, differentiated training experiences, direct work with producers, and participation in organizational spaces. During the coffee crisis, functions are reoriented toward meetings, diagnoses, productive recovery, and socio-environmental management. With the expansion of public programs, responsibilities of coordination, systematization, and connection between community knowledge and institutional requirements are incorporated.

The main differences are found in territorial origin, pathway into professional work, moment of reorientation, and type of post-crisis response. The main convergences are located in the rupture with a position centered exclusively on instruction, the permanence of work with producers, the recognition of local knowledge, and the expansion of mediation tasks. Consequently, the cases show different configurations of a common trajectory of functional expansion, conditioned by productive events, organizational experiences, and new institutional demands.

LIMITATIONS

The block compares only two cases, below the intended operational range of three to five cases.

The theme and research question were formulated from the recurrent dimensions identified in the cases and were not provided explicitly as independent inputs.

Territorial references present low geographic specificity and do not allow comparison of concrete localities.

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. *Cureus J AI Augment Res* 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

Several professional transformations have dates derived from the narrative sequence, but not exact moments of beginning or consolidation.

The correspondences between the productive crisis and individual trajectories do not allow causality to be established.

Lexical frequencies do not by themselves distinguish the narrative or evaluative function of each term.

The associations are restricted to the actors, events, resources, and relations explicitly structured in ENT_01 and ENT_02.

The findings are partial and do not allow generalization beyond the analyzed cases.

METHODOLOGY

This study adopts an Orchestrated Distant Reading approach for the analysis of life histories, treating the text as a source of data that can be structured and compared. Unlike traditional intensive reading, this approach transforms discourse into observable representations-lexical frequencies, n-grams, co-occurrences, temporal sequences, actor-event relations, and narrative profiles-following principles from statistical language processing and corpus linguistics. The notion of distant reading is used as an operational principle: it introduces analytical distance by shifting part of the analysis from the immediate interpretation of individual cases toward the observation of patterns across structured discursive instances.

The orchestrated character of the method refers to the functional segmentation of the analytical process into specialized public GPT-based instructional modules, each with delimited and controlled tasks. The workflow includes: (1) transcript cleaning; (2) structured extraction of life histories; (3) project timeline generation; (4) temporal alignment; (5) cross-case block analysis; and (6) optional inter-block synthesis. Each module receives a specific input, produces an intermediate output, and remains subject to human verification before the next stage begins. The identification of lexical associations through co-occurrence windows, together with corpus normalization, supports consistency in data treatment.

The method explicitly requires the intervention of a human researcher at each stage of the process. The researcher validates inputs, reviews structured outputs, and controls the coherence of the workflow. This responds to documented limitations in the use of large language models, particularly the risks of unsupported generation. In this sense, Orchestrated Distant Reading does not replace interpretation. It organizes evidence in a controlled, traceable, and comparable way for subsequent analysis. The name of the method reflects the combination of distant reading and modular orchestration, oriented toward maximizing traceability, comparability, and epistemological control.

REFERENCES

- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots. Proceedings of FAccT.
- Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. Computational Linguistics, 16(1), 22-29.

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. Cureus J AI Augment Res 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). Text as Data: A New Framework for Machine Learning and the Social Sciences. Princeton University Press.

Manning, C. D., & Schütze, H. (1999). Foundations of Statistical Natural Language Processing. MIT Press.

McEnery, T., & Hardie, A. (2012). Corpus Linguistics: Method, Theory and Practice. Cambridge University Press.

Moretti, F. (2013). Distant Reading. Verso.

Source link: <https://chatgpt.com/share/6a555ad7-c578-83e8-bffd-5582bbc85676>

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Joao Gabriel Rodrigues Telles Almeida

Acquisition, analysis, or interpretation of data: Joao Gabriel Rodrigues Telles Almeida, Adriana Alicia Quiroga Carapia, Eliezer Fernando Pérez Pérez, Jesús Geovani Alcázar Sánchez, Lorena Soto-Pinto

Drafting of the manuscript: Joao Gabriel Rodrigues Telles Almeida

Supervision: Joao Gabriel Rodrigues Telles Almeida, Adriana Alicia Quiroga Carapia

Critical review of the manuscript for important intellectual content: Adriana Alicia Quiroga Carapia, Eliezer Fernando Pérez Pérez, Jesús Geovani Alcázar Sánchez, Lorena Soto-Pinto

Disclosures

Human subjects: Consent was obtained from all participants in this study.

Animal subjects: All authors have confirmed that this study did not involve animal subjects or tissue.

Conflicts of interest: In compliance with [the ICMJE uniform disclosure form](#), all authors declare the following:

- **Payment/services info:** Yes. The submitted work is part of the project Life Trajectories and Social Programs in Coffee-Growing Territories of Chiapas: The Case of Sembrando Vida and PRONAI, developed at El Colegio de la Frontera Sur (ECOSUR). The project receives funding from the Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI). This support is related to the broader research project under which the study was conducted. SECIHTI did not participate in the study design, data collection, analysis, interpretation of results, manuscript preparation, or the decision to submit the article for publication.
- **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work.
- **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

AI Disclosure Statement

The authors used ChatGPT (OpenAI; GPT-4.4; GPT-5.5) during manuscript development and methodological reconstruction to support the organization, drafting, and revision of the title, abstract, keywords, introduction, methods, results, and conclusion. Public custom GPTs were also used as instructional modules during the analytical workflow

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. Cureus J AI Augment Res 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

described in the manuscript, including transcript cleaning, structured extraction, temporal baseline construction, temporal alignment, narrative profiling, cross-case comparison, and controlled synthesis. All AI-assisted outputs were manually reviewed and verified by the authors against the original interview materials, intermediate analytical outputs, and relevant methodological literature. Outputs were revised, corrected, or discarded when they introduced unsupported inferences, imprecise formulations, or inconsistencies with the empirical material. The authors take full responsibility for the accuracy, integrity, and originality of the content.

Data Availability Statements

The datasets (and/or code) supporting this study are available from the corresponding author upon reasonable request.

Acknowledgements

This article is part of the project Life Trajectories and Social Programs in Coffee-Growing Territories of Chiapas: The Case of Sembrando Vida and PRONAIL, developed at El Colegio de la Frontera Sur (ECOSUR). The project is supported by the Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI) and focuses on a qualitative analysis, based on life histories, of the subjective, cultural, and relational impacts of public policies in coffee-growing territories of Chiapas.

References

1. Weber M: [The Methodology of the Social Sciences](#). Free Press, 1949.
2. Schutz A: [The Phenomenology of the Social World](#). Northwestern University Press, 1967.
3. Geertz C: [The Interpretation of Cultures](#). Basic Books, 1973.
4. Goffman E: [The Presentation of Self in Everyday Life](#). Anchor Books, 1959.
5. Thomas WI, Znaniecki F: [The Polish Peasant in Europe and America: Monograph of an Immigrant Group \(Vols. 1-5\)](#). University of Chicago Press, 1918.
6. Ferrarotti F: [On the Science of Uncertainty: The Biographical Method in Social Research](#). Lexington Books, 2005.
7. Ferrarotti F: [Biography and the Social Sciences](#). Social Research. 1983, 57-80.
8. Bourdieu P: [The biographical illusion](#). Biography in Theory: Key Texts with Commentarie. Hemecker W, Saunders E (ed): De Gruyter, Berlin; 2017. 210-217. [10.1515/9783110516678-036](#)
9. Clot Y: [La otra ilusión biográfica \[The other biographical illusion\]](#). Acta Sociológica. 2011, 56:129-134. [10.22201/fcpys.24484938e.2011.56.29461](#)
10. Denzin NK: [The Research Act: A Theoretical Introduction to Sociological Methods](#). McGraw-Hill, 1978.
11. Lincoln YS, Guba EG: [Naturalistic Inquiry](#). SAGE, 1985.
12. Finlay L: ["Outing" the Researcher: The Provenance, Process, and Practice of Reflexivity](#). Qualitative Health Research. 2002, 12:531-545. [10.1177/104973202129120052](#)
13. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA: [Standards for reporting qualitative research: A synthesis of recommendations](#). Academic Medicine. 2014, 89:1245-1251. [10.1097/acm.0000000000000388](#)
14. Tong A, Sainsbury P, Craig J: [Consolidated criteria for reporting qualitative research \(COREQ\): a 32-item checklist for interviews and focus groups](#). International Journal for Quality in Health Care. 2007, 19:349-357. [10.1093/intqhc/mzm042](#)
15. Grimmer J, Stewart BM: [Text as data: The promise and pitfalls of automatic content analysis methods for political texts](#). Political Analysis. 2013, 21:267-297. [10.1093/pan/mps028](#)
16. McEnery T, Hardie A: [Corpus Linguistics: Method, Theory and Practice](#). Cambridge University Press, 2012. [10.1017/CBO9780511981395](#)
17. Moretti F: [Distant Reading](#). Verso, 2013.
18. Jockers ML: [Macroanalysis: Digital Methods and Literary History](#). University of Illinois Press, 2013. [10.5406/illinois/9780252037528.001.0001](#)

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. Cureus J AI Augment Res 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>

19. Varela FJ: [Neurophenomenology: A methodological remedy for the hard problem](#). Journal of Consciousness Studies. 1996, 3:330-349.
20. Gilardi F, Alizadeh M, Kubli M: [ChatGPT outperforms crowd workers for text-annotation tasks](#). Proceedings of the National Academy of Sciences of the United States of America. 2023, 120:e2305016120. [10.1073/pnas.2305016120](#)
21. Hitch D: [Artificial intelligence augmented qualitative analysis: The way of the future?](#). Qualitative Health Research. 2024, 34:595-606. [10.1177/10497323231217392](#)
22. Morgan DL: [Exploring the use of artificial intelligence for qualitative data analysis: The case of ChatGPT](#). International Journal of Qualitative Methods. 2023, 22:10.1177/16094069231211248
23. Tai RH, Bentley LR, Xia X, Sitt JM, Fankhauser SC, Chicas-Mosier AM, Monteith BG: [An examination of the use of large language models to aid analysis of textual data](#). International Journal of Qualitative Methods. 2024, 23:10.1177/16094069241231168
24. Bender EM, Gebru T, McMillan-Major A, Shmitchell S: [On the dangers of stochastic parrots: Can language models be too big?](#). FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021, 610-623. [10.1145/3442188.3445922](#)
25. Park SH, Suh CH, Lee JH, Tejani AS, You SC, Kahn CE, Moy L: [Minimum reporting items for clear evaluation of accuracy reports of large language models in healthcare \(MI-CLEAR-LLM\): 2025 updates](#). Korean Journal of Radiology. 2025, 26:1123-1132. [10.3348/kjr.2025.1522](#)
26. Varela FJ, Thompson E, Rosch E: [The Embodied Mind: Cognitive Science and Human Experience](#). MIT Press, 1991.
27. Rodrigues Telles Almeida JG, Quiroga Carapia AA, Soto Pinto ML, et al.: [Trayectorias de vida y programas sociales en territorios cafetaleros de Chiapas: el caso de Sembrando Vida y PRONAI: Informe narrativo analítico de dos historias de vida en relación con trayectorias de políticas públicas \[Life trajectories and social programs in coffee-growing territories of Chiapas: the case of Sembrando Vida and PRONAI: analytical narrative report of two life histories in relation to public policy trajectories\]](#). Ecosur, México; 2026.
28. Ferrarotti F: [Las historias de vida como método \[Life histories as a method\]](#). Acta Sociológica. 2011, 56:95-119.
29. Ferrarotti F: [Storia e storie di vita \[History and life stories\]](#). Laterza, 1981.
30. Feixa C: [La imaginación autobiográfica \[The autobiographical imagination\]](#). Perifèria: revista de recerca i formació en antropologia. 2006, 5:1-44. [10.5565/rev/periferia.159](#)
31. OpenAI: [GPTs in ChatGPT](#). OpenAI Help Center. 2026a,
32. OpenAI: [Creating a GPT](#). OpenAI Help Center. 2026b,
33. Church KW, Hanks P: [Word association norms, mutual information, and lexicography](#). Computational Linguistics. 1990, 16:22-29.
34. Dai SC, Xiong A, Ku LW: [LLM-in-the-loop: Leveraging large language model for thematic analysis](#). Association for Computational Linguistics. Bouamor H, Pino J, Bali K (ed): Association for Computational Linguistics, Singapore; 2023. 9993-10001. [10.18653/v1/2023.findings-emnlp.669](#)
35. Gao J, Guo Y, Lim G, et al.: [CollabCoder: A lower-barrier, rigorous workflow for inductive collaborative qualitative analysis with large language models](#). CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. 2024, 1-29. [10.1145/3613904.3642002](#)
36. Frieze S: [From coding to conversation: A new methodological framework for AI-assisted qualitative analysis](#). Qualitative Inquiry. 2026, [10.1177/10778004251412871](#)
37. Xu H, Yi S, Lim T, et al.: [TAMA: A human-AI collaborative thematic analysis framework using multi-agent LLMs for clinical interviews](#). ACM Transactions on Computing for Healthcare. 2026, [10.1145/3828752](#)

How to cite this article:

Rodrigues Telles Almeida J, Pérez Pérez E, Alicia Quiroga Carapia A, et al. (September 04, 2026) Orchestrated Distant Reading: A Generative Pre-trained Transformer (GPT)-Based Modular Workflow for Traceable Life-History Analysis. Cureus J AI Augment Res 1 : es44568-026-00263-6. DOI <https://doi.org/10.7759/s44568-026-00263-6>