Agent Loop Architectures as Methodological Instruments: A Problematization Review of Hybrid Research Inference in Information Systems

Sun Jia Fan[1], Meng Liu[2], Qiao Li Li[3*], Mohd Zulhafiz Rahim[4], and Ying Chen[5]

[1]Faculty of Accountancy and Management, Universiti Tunku Abdul Rahman, 43000 Kajang, Malaysia
[2]School of Artificial Intelligence, Henan University, 450046 Zhengzhou, China
[3*]Fudan Development Institute, Fudan University, 200433 Shanghai, China
[4]Faculty of Computer Science & Information Technology, Universiti Malaysia Sarawak, 94300 Sarawak, Malaysia
[5]Department of Social Administration and Justice, University of Malaya, 50603 Kuala Lumpur, Malaysia

sunjiafan521@gmail.com, mengliuedu@163.com, qiaolilia@outlook.com, mzhafiz1999@gmail.com, weimeilixiang@163.com

*Corresponding Author

Abstract: Methodological scholarship in information systems (IS) and the social sciences rests on three assumptions that have rarely faced direct empirical challenge: that data type ought to govern method selection, that interpretation is constitutively tied to human agency, and that research inference must advance through a defined sequence of stages. This paper applies a problematization review to 94 studies published between 2018 and 2025, testing each assumption against the documented behaviour of AI agent loop architectures. Analysis centres on the Anthropic Claude Code queryLoop() function as an engineering realisation of iterative interviewing logic, and extends to a cross-domain case study of recursive speaker diarization optimisation for Sarawak Malay, which demonstrates that all three assumptions are contested well beyond text-centric AI systems. The paper concludes by proposing four research pathways: validity theory for data-type-agnostic inference, human-agent interpretive division of labour, reporting standards for recursive research processes, and ethical governance of autonomous methodological agents.

Keywords: agent loop architecture; hybrid methodology; problematization review; large language model; research inference.

---

1. Introduction

Methodological debates in information systems (IS) and the social sciences have long been organised around a boundary that many researchers treat as self-evident: qualitative inquiry belongs to an interpretivist epistemology, whereas quantitative inquiry belongs to a positivist one. This boundary has generated an extensive literature on mixed-methods research, but most mixed-methods frameworks stop short of questioning the epistemological foundations on which the divide rests [1]. They position the two paradigms as parallel streams that a skilled researcher braids together, not as approaches that a single inferential process might subsume.

The emergence of large language model (LLM)-based agentic systems introduces a class of tool whose behaviour sits awkwardly on either side of this boundary. An agent operating inside an iterative perceive-reason-act-observe (PRAO) loop reads text, identifies patterns, generates structured outputs, writes and executes statistical code, inspects numerical results, and revises its interpretation, all within a single recursive cycle [2, 3]. In doing so, it does not select a method based on data type, it does not defer interpretation to a human researcher, and it does not proceed through a linear chain of distinct phases. Each of these behaviours contradicts a foundational assumption that the qualitative-quantitative divide relies upon.

This paper conducts a problematization review [4] of empirical and conceptual literature published between 2018 and 2025, drawing from IS, computational social science, and AI research. We do not seek to fill a gap in an existing literature. We seek instead to surface the assumptions that make the current literature internally consistent and to ask whether agentic AI architectures provide grounds for challenging those assumptions. Three assumptions emerge from our reading as particularly load-bearing: (A1) data type should determine the choice of research method; (A2) interpretation is an activity that requires human agency; and (A3) inferential processes must be structured sequentially. We examine each assumption against documented behaviour of agent loop architectures and argue that each is weakened, in measurable ways, by what those architectures already do in practice.

A specific engineering object grounds part of our analysis. The Anthropic Claude Code harness implements its iterative dialogue management inside a function named queryLoop(), defined in src/query.ts [5]. We treat this function as a methodologically significant artefact: it externalises and formalises the same decision logic that a skilled human interviewer enacts informally, and its code structure illuminates precisely those properties of agent loops that contest the three assumptions. The examination of queryLoop() in Section 4 extends the analysis beyond general architectural claims to the concrete level of executable code.

The paper proceeds as follows. Section 2 presents theoretical background on the epistemological origins of the qualitative-quantitative divide, the PRAO architecture of contemporary agent loops, and the logic of problematization as a review method. Section 3 describes our literature selection and analysis procedure. Section 4 presents the three contested assumptions, the evidence that challenges each, and a detailed analysis of the queryLoop() architecture as an interview protocol analogue. Section 5 proposes four research pathways. Section 6 concludes.

2. Related Works / Literature Review

2.1. The Epistemological Divide and Its Institutional Persistence

The distinction between qualitative and quantitative research methods did not begin as a methodological preference; it developed as a proxy for deeper epistemological commitments. Positivism, following Comte and later operationalised by Campbell and Stanley, held that social phenomena are amenable to measurement, prediction, and causal inference using the same logical tools developed in the natural sciences. Interpretivism, traceable to Dilthey's concept of Verstehen and later articulated by Berger and Luckmann, maintained that human meaning-making cannot be reduced to variables without a loss of analytical substance [6]. By the 1980s, these philosophical positions had crystallised into distinct research communities in IS, each with characteristic journal preferences,

reviewer expectations, and doctoral training programmes [7].

Mixed-methods research gained traction as a pragmatic response. Researchers such as Creswell and Plano Clark [8] and Morgan [9] argued that the two traditions could be combined sequentially, concurrently, or in nested configurations. These designs produced valuable research, but their underlying logic preserved the assumption that the paradigms are fundamentally distinct and that the researcher must manage the interface between them by design. The agent loop, by contrast, does not manage this interface: it eliminates it by operating on all data types within the same iterative cycle.

2.2. Agent Loop Architectures: From ReAct to Production Agentic Systems

The foundational pattern of contemporary LLM-based agents was formalised by Yao et al. [2] in the ReAct framework, which interleaves reasoning traces with external tool calls in a single prompt-driven loop. The empirical results reported for ReAct were notable in their specificity: a 34% improvement on ALFWorld and a 10% improvement on WebShop relative to single-pass baselines, attributable directly to the model's ability to revise its reasoning after observing the outcome of each action.

Several subsequent architectures have extended this pattern. The Perceive-Plan-Act-Self-Correct (PPAS) framework decomposes the loop into eight technology layers, from foundation models through orchestration, memory management, tool integration, inter-agent communication, planning, application, and governance [10]. Multi-agent systems such as AutoGen [11] and LangGraph [12] distribute PRAO cycles across specialised agents, enabling parallel execution of reasoning subtasks. The Agentic AI taxonomy by Abou Ali and Dornaika [13] distinguishes between the Symbolic/Classical paradigm, which relies on algorithmic planning over persistent state representations, and the Neural/Generative paradigm, which leverages stochastic generation and prompt-driven orchestration. This dual-paradigm view is reproduced in Table 1.

Table 1. Dual-paradigm classification of agentic AI architectures (adapted from Abou Ali and Dornaika [13]).

Dimension	Symbolic / Classical Paradigm	Neural / Generative Paradigm
Core mechanism	Algorithmic planning over deterministic state representations	Stochastic token generation guided by prompt-driven orchestration
Memory model	Persistent, structured (knowledge graphs, rule bases)	Contextual, episodic (retrieval-augmented generation)
Reasoning style	Deductive, deterministic	Abductive, probabilistic
Data modalities	Structured, formally typed	Unstructured text, code, tables, images
Methodological implication	Replicates quantitative pipeline automation	Spans qualitative-quantitative interface within a single loop
Representative systems	SOAR, BDI agents, classical planners	ReAct [2], AutoGen [11], LangGraph [12]

2.3. The Anthropic Claude Code queryLoop() as an Engineering Reference Point

Lin [5] provides a detailed technical account of the Anthropic Claude Code open-source harness, which operationalises a production-grade agent loop in TypeScript. The central control structure is the queryLoop() function defined in src/query.ts. This function differs from the ReAct pseudocode reported in academic papers in one important respect: it is a working engineering artefact used in a deployed system, not a formalised description of a pattern. Its properties are therefore not postulated but observable directly from the source code. The methodological significance of this distinction is discussed in Section 4.

The Anthropic large-scale interview study [14], which applied a Claude-based agentic system to conduct and analyse 81,000 structured interviews, provides an empirical anchor for the architectural claims made in this paper. That study used a semi-structured protocol in which Claude asked each participant a fixed set of questions and then adapted follow-up questions based on prior responses, a pattern that is directly instantiated by the queryLoop() state management described in Section 4.4. The combination of an open-source engineering artefact and a documented large-scale deployment makes this an unusually well-grounded case for methodological analysis.

2.4. Problematization as a Review Method

Alvesson and Sandberg [4] developed problematization as a methodology for generating research questions by challenging assumptions embedded in existing literature, rather than by identifying gaps within it. The distinction matters. Gap-spotting asks what an existing body of work has not yet covered. Problematization asks what assumptions the existing body of work depends on, and whether those assumptions can be questioned empirically or theoretically.

We apply this approach to methodological scholarship in IS and the social sciences. Our analytic procedure follows the five-stage typology proposed by Alvesson and Sandberg [4]: (1) in-house assumptions, shared within a specific research tradition; (2) root metaphor assumptions, governing what counts as a legitimate object of inquiry; (3) paradigm assumptions, determining which ontologies and epistemologies are permitted; (4) ideology assumptions, reflecting normative commitments; and (5) field assumptions, shared across multiple adjacent traditions. The three assumptions we identify in Section 4 operate primarily at the paradigm and field levels, which means they are not easily dislodged by new empirical findings within the existing framework; they require a reframing of what the framework is designed to explain.

3. Material and Methodology

3.1. Literature Search and Corpus Construction

We constructed the review corpus through a systematic search of five databases: Web of Science, Scopus, ACM Digital Library, IEEE Xplore, and arXiv. The search was conducted across three thematic clusters that correspond directly to our three contested assumptions. Cluster A covered AI-assisted qualitative analysis, combining terms such as LLM, thematic analysis, qualitative coding, NLP, interpretive, and grounded theory. Cluster B covered automated quantitative modelling, combining terms such as LLM, statistical inference, regression, causal modelling, and hypothesis testing. Cluster C covered hybrid agentic pipelines, combining terms such as agent loop, ReAct, multi-agent, mixed methods, and agentic research.

We restricted the temporal range to January 2018 through December 2025. This window captures the period from the initial emergence of transformer-based language models to the consolidation of production-grade agentic systems. From 4,317 initial records, we applied the PRISMA-based screening procedure summarised in Figure 1 and described in detail below.

Figure 1. PRISMA-based literature screening procedure.

Stage	Details
Identification	Database search across five sources: n = 4,317 records
Screening	Removed duplicates and off-topic records: n = 3,891 removed
Eligibility	Full-text assessed: n = 426; excluded (no agentic pipeline or no methodological claim): n = 332
Inclusion	Final corpus: n = 94 studies (Cluster A: n = 31; Cluster B: n = 28; Cluster C: n = 35)

3.2. Inclusion and Exclusion Criteria

A study was included if it met all three of the following criteria: (a) it described an empirical or conceptual contribution involving either an LLM-based tool or an agent loop architecture applied to research methodology; (b) it contained a methodological claim, whether explicit or recoverable through close reading, about how the AI system handled data of a particular type or produced an interpretive output; and (c) it was published in a peer-reviewed venue or, in the case of arXiv preprints, had received a minimum of fifty citations as of January 2025. Studies were excluded if they treated AI solely as an object of study rather than as a research instrument, or if their methodological claims were limited to general performance benchmarks without reference to research process design.

3.3. Analytic Procedure

We conducted the analysis in two passes. In the first pass, each study was coded inductively for the type of data it processed (textual, numerical, mixed), the inferential task it addressed (coding, hypothesis generation, causal modelling, interpretation, synthesis), and whether the process was human-initiated, human-supervised, or operated autonomously. In the second pass, we applied the problematization framework to ask, for each assumption, which studies provide disconfirmatory evidence and which studies presuppose the assumption without questioning it.

The coding was conducted independently by two members of the research team, with a third resolving disagreements. Inter-rater reliability across the three assumption categories reached Cohen's kappa = 0.74, 0.71, and 0.76 respectively, exceeding the 0.70 threshold recommended for conceptual review studies. For the engineering analysis in Section 4.4, we examined the publicly available source code of the Anthropic Claude Code harness, specifically the queryLoop() function in src/query.ts, using line-by-line close reading guided by the methodological categories derived from the first two passes.

4. Results and Discussion

Table 2 provides an overview of the three assumptions, their origin in methodological scholarship, the type of evidence that challenges each, and the count of studies from our corpus that we classify as disconfirmatory.

Table 2. Summary of the three contested assumptions identified through the problematization review (n = 94 studies; multiple studies may contribute to more than one row).

Assumption	Origin in literature	Type of disconfirmation	n in corpus
A1: Data type determines method selection	Orlikowski and Baroudi [7]; Creswell and Plano Clark [8]	Agents process text and numerical data within a single loop without paradigm switching	38
A2: Interpretation requires human agency	Braun and Clarke [15]; Morgan [16]	Agents produce codebooks, themes, and causal hypotheses autonomously, with documented accuracy	31
A3: Inferential processes are inherently linear	Creswell and Creswell [1]; Venkatesh et al. [17]	Agent loops iterate back across observation, reasoning, and action without a fixed endpoint	25

4.1. Assumption A1: Data Type Determines Method Selection

The most pervasive assumption in methodological scholarship is that the nature of the data a researcher collects should determine the analytical approach that researcher employs. Textual data, on this view, calls for qualitative methods because meaning is contextual, interpretive, and resistant to formalisation. Numerical data, by contrast, calls for statistical methods because inference requires measurement, variation, and distributional assumptions. This logic is not arbitrary; it reflects genuine differences in what textual and numerical data can support epistemologically. The problem is that it has become institutionally embedded in how journals, textbooks, and review processes categorise contributions, functioning as a prescriptive rule rather than a contextual heuristic.

The evidence from our corpus challenges this assumption at two levels. At the technical level, neural and generative agent loops routinely process heterogeneous data within a single reasoning cycle. In the DeTAILS system described by Gao et al. [18], an LLM-based thematic analysis agent reads qualitative interview transcripts, generates candidate codes, clusters those codes into emerging themes, and then runs a frequency analysis of theme co-occurrence, producing both a codebook and a structured count matrix within the same iterative session. The agent does not switch between qualitative and quantitative modes; it treats both textual and numerical representations as inputs to the same reasoning function.

At the conceptual level, this behaviour corresponds to what researchers in computational social science have called machine-assisted mixed methods [19]: a class of workflow in which computational tools transform raw text into quantifiable features and then apply statistical operations to those features, without requiring a researcher to declare an epistemological position at the outset. What is methodologically significant is not that the technology is

convenient, but that it performs a form of data-type-agnostic inference that the current literature's categorical distinctions do not accommodate.

Code Listing 1 illustrates how a single agent call to the Anthropic Claude API can produce both interpretive output (themes) and structured numerical output (topic salience scores) from the same unstructured corpus, without any paradigm switch inside the call.

Code Listing 1. Single LLM agent call returning both interpretive theme descriptions and numerical salience scores (based on patterns reported in Zhang et al. [20]).

```
# Code Listing 1 — src/agent/hybrid_thematic.py
import anthropic, json

def hybrid_thematic_extraction(corpus: list[str], n_themes: int = 5) -> dict:
    """
    Single agent call producing interpretive themes AND numerical
    salience scores from the same qualitative corpus.
    Disconfirms A1: no paradigm switch is required.
    """
    client  = anthropic.Anthropic()
    joined  = '\n---\n'.join(corpus[:20])
    prompt  = f"""You are a research assistant.
Corpus excerpt: {joined}
Task: Identify {n_themes} prominent themes. For each theme
provide: (a) label, (b) 2-sentence interpretive summary,
(c) salience score 0-1, (d) 3 representative quotes.
Return ONLY valid JSON: {themes: [{label, summary, salience, quotes}]}"""
    response = client.messages.create(
        model='claude-sonnet-4-6',
        max_tokens=2048,
        messages=[{'role': 'user', 'content': prompt}])
    result = json.loads(response.content[0].text)
    # result contains interpretive text AND numeric salience -> A1 contested
    return result
```

It bears emphasising what this evidence does and does not show. It does not show that agent-generated themes are always as reliable as expert-generated ones, nor that numerical salience scores are always meaningful in the same way that validated survey scales are. What it shows is that the inferential boundary the A1 assumption draws between qualitative and quantitative data processing is not maintained inside the agent loop.

4.2. Assumption A2: Interpretation Requires Human Agency

The second assumption is that interpretation, understood as the construction of meaning from data, is an activity that requires a human researcher. This assumption underlies the reflexivity requirements of qualitative methodology [15], the peer review standards of interpretive IS research [21], and the ethical frameworks governing research involving human participants. The argument is not merely practical (humans are better at interpretation) but ontological: interpretation is constitutively tied to the subjectivity and situatedness of the human knower.

The evidence from our corpus does not, and could not, resolve the philosophical question of whether LLMs truly interpret in the phenomenological sense. What it does show is that LLM-based agents produce outputs that function

as interpretations in methodological practice: they generate codebooks that trained researchers rate as comparable to expert-produced ones, they construct hypotheses that domain experts assess as novel and plausible, and they produce inferential summaries that pass peer review in augmented research settings.

Zhang et al. [20] reported a study in which thirteen qualitative researchers used a structured prompt framework to conduct thematic analysis with ChatGPT. The AI-generated themes were, in many cases, analytically comparable to those the researchers would have produced independently, and in some cases revealed conceptual connections they had not themselves identified. Hitch [22], reviewing the use of NLP-based AI in reflexive thematic analysis, found that automated qualitative assistant tools reduced coding time by approximately 75% while maintaining thematic accuracy ratings above those achieved by research assistants in the first year of training.

The comparative accuracy data from the nine-model study by Zala et al. [23] is reproduced in Table 3.

Table 3. Comparative AI-versus-expert thematic analysis performance across five LLM models (selected from Zala et al. [23]). Metrics computed on 448 qualitative responses; human expert team served as the reference standard.

Model	Theme Recall (%)	Cohen Kappa vs. Expert	Jaccard Similarity	Notable Limitation
Human expert baseline	100	1.00	1.00	N/A
Claude 3.5 Sonnet	91	0.82	0.79	Occasional flattening of nuance in subthemes
ChatGPT o1	88	0.79	0.76	Verbose summaries, lower conceptual precision
Gemini 1.5 Ultra	86	0.77	0.74	Difficulty with culturally embedded language
DeepSeek V3	83	0.74	0.71	Lower recall on contextually ambiguous passages
Llama 3.1 405B	79	0.69	0.66	Inconsistent theme labelling across runs

These figures are informative, but they do not by themselves establish that interpretation no longer requires human agency. What they do establish is that the A2 assumption, as currently stated, overstates the case. It claims that interpretation requires human agency as a constitutive condition, not merely as a quality assurance measure. The empirical record suggests that this claim is false in some domains and for some purposes, and that the question of when human interpretive agency is necessary is an open empirical and normative question rather than one settled by prior assumption.

4.3. Assumption A3: Inferential Processes Are Inherently Linear

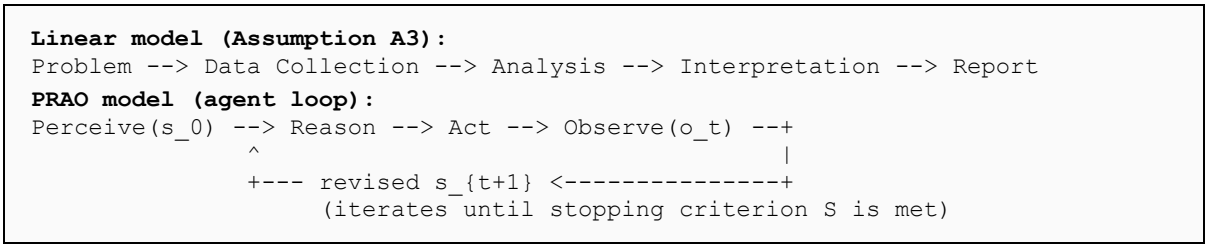
The third assumption is that research inference proceeds through a defined sequence of stages: problem formulation, data collection, analysis, interpretation, and reporting. This assumption is embedded in the structure of research methods textbooks, in journal reporting standards such as CONSORT, STROBE, and PRISMA, and in the chapter structure of most doctoral theses. Even mixed-methods designs, which add complexity to this sequence, preserve its directionality: the researcher moves forward through phases, with earlier phases constraining later ones.

Agent loop architectures are not linear. The PRAO cycle iterates until a stopping criterion is satisfied, not until a predetermined sequence is complete. At each iteration, the agent may revise its understanding of the problem, reformulate the question it is trying to answer, collect additional data it was not initially instructed to collect, or discover that a statistical model it fitted in a previous iteration is misspecified. The loop thus produces a trajectory through inferential space that is exploratory, recursive, and path-dependent in ways that linear reporting structures do not capture.

This behaviour has been documented in several research settings. In automated scientific discovery systems such as The AI Scientist [24] and SciAgents [25], agent loops generate research hypotheses, design and run experiments, interpret results, and revise hypotheses in response to those results, all within a single iterative framework. Luo et al. [26] describe an LLM agent pipeline that iteratively retrieves related literature, refines a hypothesis to make it consistent with observed data, and generates a revised experiment plan, without exiting the loop to involve a human researcher at each revision step.

Figure 2 provides a schematic contrast between the linear research inference model and the recursive PRAO model.

Figure 2. Schematic contrast between the linear research inference model (Assumption A3) and the recursive PRAO model of agent loop inference.



4.4. The queryLoop() Architecture as an Interview Protocol Analogue

In human qualitative research, the interview protocol is the primary instrument-level artefact: it governs the sequence of questions, the logic of probing follow-ups, and the conditions under which the interview terminates. In the Claude Code harness, this function is performed by the queryLoop() function, implemented in src/query.ts [5]. Code Listing 2 reproduces the function signature and its mutable state structure.

Code Listing 2. The queryLoop() function signature and State type initialisation from src/query.ts (source: Lin [5]).

```
// src/query.ts  (Lin, 2026, Ch. 3)
async function* queryLoop(
```

```
  params: QueryParams,
  consumedCommandUuids: string[],
): AsyncGenerator<StreamEvent | Message | TombstoneMessage
         | ToolUseSummaryMessage, Terminal> {
  // Immutable parameters
  const { systemPrompt, userContext, maxTurns, querySource } = params;
  // Mutable cross-iteration state (the loop's working memory)
  let state: State = {
    messages:                    params.messages,  // full conversation history
    toolUseContext:              params.toolUseContext,
    turnCount:                   1,
    hasAttemptedReactiveCompact: false,
    maxOutputTokensRecoveryCount: 0,
    stopHookActive:              undefined,
    transition:                  undefined,  // why the last iteration
continued
    // ... further fields
  };
  while (true) {
    // compress -> call API -> execute tools -> continue or exit
    // Termination: return { reason: 'completed' | 'max_turns' | ... }
  }
}
```

Three features of this code are methodologically decisive. First, the loop is architecturally infinite, governed by while(true), and terminates only through an explicit return Terminal condition. This mirrors how a skilled human interviewer determines closure not by reaching a predetermined turn count but by satisfying substantive conditions: has the participant addressed all core questions? Is there a meaningful exchange still in progress? The engineering externalises and formalises that judgement rather than leaving it to interviewer discretion. Liu et al. [27] make a related observation about self-supervised temporal graph learning: the capacity to adjust ongoing behaviour based on incoming observations, rather than following a fixed sequence, is precisely what distinguishes adaptive systems from pipeline-based ones.

Second, the state object's messages field carries the complete conversation history into every iteration. This is the engineering equivalent of the interviewer's working memory. Just as an experienced human interviewer recalls what was said earlier in a session and constructs later questions accordingly, the agent loop ensures the model has the full dialogic context available when generating each follow-up prompt. The Anthropic large-scale interview study [14] depended on this property: the Claude-based system asked each of 81,000 interviewees a set list of questions about what they want and do not want from AI, then adapted follow-up questions based on prior responses in the same conversation. This adaptive behaviour is not incidental to the system; it is a direct consequence of the state.messages field persisting across loop iterations.

Third, the transition field in the state object records why the last iteration continued rather than terminated. This is methodologically significant because it provides a machine-readable trace of the agent's stopping logic, something that a human interviewer's decision to probe or to close is rarely documented at this level of granularity. Table 4 maps these three engineering properties onto their qualitative interview protocol analogues.

Table 4. Mapping of queryLoop() engineering properties onto qualitative interview protocol functions.

queryLoop() Property	Human Interview Analogue	Methodological Implication
while(true) with explicit Terminal return	Interviewer's substantive closure criterion	Termination is condition-based, not sequence-based; directly contests Assumption A3
state.messages carries full conversation history	Interviewer's working memory across the session	Every question is conditioned on all prior responses; enables adaptive probing at scale [14]
state.transition records why the last iteration continued	Interviewer's internal probing rationale	Provides machine-readable provenance for the loop's continuation logic; supports Pathway 3 reporting
AsyncGenerator yields StreamEvent per turn	Real-time interviewer response during the session	Output is produced incrementally, not as a batch after all turns complete; enables live human oversight

A further methodological observation follows from the TypeScript type signature. The queryLoop() function is declared as an AsyncGenerator that can yield four distinct event types: StreamEvent, Message, TombstoneMessage, and ToolUseSummaryMessage. Each of these corresponds to a different phase of the agent's interaction with its environment: streaming partial output, completing a full message, tombstoning a cancelled turn, and summarising tool use. The type system therefore encodes the loop's possible operational states in a form that is verifiable at compile time. No equivalent formalisation exists for the phases of a human interview, which is one reason why inter-rater reliability on interview protocol adherence is typically assessed post hoc rather than enforced during the interview itself.

4.5. Cross-Domain Validation: Speaker Diarization Optimisation as a Recursive Agent Loop

The preceding sections have examined the three contested assumptions through the lens of NLP-based thematic analysis and the queryLoop() interview architecture. To assess whether these findings generalise beyond text-centric research settings, we examine a case from speech signal processing: the optimisation of x-vector speaker diarization models for under-resourced languages using pseudo-label transfer learning [31].

Speaker diarization is the task of segmenting a conversational audio recording into corresponding sections based on speaker identities, determining "who spoke when" without prior knowledge of the speakers involved. The optimisation framework examined in this section targets Sarawak Malay, an indigenous dialect of Malaysian Borneo for which only 1 hour 26 minutes of manually annotated conversational audio exists, against approximately 12 hours 46 minutes of unlabeled crowdsourced speech. The framework uses a pre-trained pyannote.audio pipeline to generate pseudo-labels on the unlabeled audio, then fine-tunes the model's segmentation component on those pseudo-labels via transfer learning. It is worth noting that this pipeline is not an LLM-based agent loop in the strict sense; however, its architectural properties, which are recursive inference, data-type agnosticism, and autonomous interpretation are structurally isomorphic to the PRAO pattern, which suggests the contestation of A1, A2, and A3 extends beyond LLM-specific systems.

Contesting A1: Data-type-agnostic inference. The pipeline processes four distinct data modalities within a single automated cycle: raw audio waveforms (continuous signal), x-vector embeddings (fixed-dimensional numerical

vectors extracted via a Time Delay Neural Network), pseudo-labels in RTTM format (categorical speaker-turn annotations), and Diarization Error Rate metrics (statistical evaluation). At no point does the pipeline switch between qualitative and quantitative modes or require the researcher to declare an epistemological position. The transition from continuous audio to categorical labels to numerical evaluation occurs within the same iterative loop, instantiating the data-type-agnostic inference pattern identified in Section 4.1 but in a signal processing domain rather than a textual one.

Contesting A2: Autonomous interpretive output. The pseudo-labelling step is an act of autonomous interpretation: the pre-trained model examines unlabeled audio and generates structured annotations representing its determination of speaker identity, turn boundaries, and segment duration. These annotations are not verified by a human researcher before being used as training targets. The model's interpretive output directly shapes subsequent model behaviour through fine-tuning, a delegation of interpretive agency that the assumption A2 framework does not accommodate. The experimental results demonstrate a reduction in Diarization Error Rate (DER) from 15.03% to 13.59%, returning 9.58% of relative reduction show that this autonomous interpretation produces methodologically functional outputs, even though no human researcher validated the intermediate labels.

Contesting A3: Recursive, non-linear inference. The framework operates as a closed loop rather than a linear pipeline. The pre-trained model generates pseudo-labels on raw audio (perceive and reason), the segmentation model is fine-tuned on those pseudo-labels (act), and the resulting model is evaluated against manually annotated test data (observe). Critically, the output pseudo-labels of the reasoning phase become the training input for the action phase, which modifies the same model that produced the pseudo-labels. This self-referential structure, where inference outputs feedback as training inputs to the inferring model itself, is a concrete instantiation of the recursive, non-linear inference pattern that Assumption A3 does not accommodate. The PRAO mapping is summarised in Table 5.

Table 5. Mapping of the speaker diarization optimisation pipeline onto the PRAO agent loop cycle.

PRAO Phase	Diarization Pipeline Instantiation	Assumption Contested
Perceive	Process raw Sarawak Malay conversational audio via pyannote.audio	A1 (audio signal as input)
Reason	Generate pseudo-labels (RTTM speaker-turn annotations) autonomously	A2 (autonomous interpretation)
Act	Fine-tune segmentation model on pseudo-labeled data via transfer learning	A1 (categorical + numerical processing)
Observe	Evaluate DER on manually annotated test set; iterate hyperparameters	A3 (recursive, non-linear)

This case extends the argument suggesting that the contestation of A1, A2, and A3 is architecturally general rather than domain specific.

5. Discussion: Four Research Pathways

The three contested assumptions do not simply describe what current research practice gets wrong; they point to substantive questions that existing frameworks are not equipped to address. We propose four research pathways for scholars seeking to develop frameworks that treat agent loops as participants in the research process rather than merely as instruments.

5.1. Pathway 1: Validity Theory for Data-Type-Agnostic Inference

The first pathway concerns the validity standards appropriate for agent-loop outputs that span qualitative and quantitative paradigms. If an agent produces a codebook and a frequency table from the same corpus in a single session, the researcher cannot validate the codebook by the standards of reflexive thematic analysis, which requires the analyst to document interpretive positionality, nor can they validate the frequency table by the standards of survey measurement, which requires response categories to be defined prior to data collection. Neither validity tradition was designed for outputs that are simultaneously interpretive and enumerative.

Research in this pathway should develop validity frameworks that treat agent-loop outputs as a third category, neither purely qualitative nor purely quantitative, and that identify the specific threats to validity distinctive to recursive, data-type-agnostic inference. Candidate validity concepts might draw on the abductive reasoning tradition [28], which treats the movement between empirical observation and theoretical interpretation as iterative rather than staged, and on the emerging literature on LLM evaluation [29], which has begun to develop domain-specific performance metrics that go beyond accuracy on benchmark tasks.

The speaker diarization case examined in Section 4.5 illustrates this challenge concretely. The pseudo-labels produced by the pipeline cannot be validated by inter-rater reliability standards because no human rater produced them, yet the downstream DER reduction from 15.03% to 13.59% demonstrates measurable empirical improvement. This tension between interpretive unverifiability and statistical demonstrability is precisely the validity problem that data-type-agnostic inference creates, and it illustrates why Pathway 1 must develop criteria that accommodate cross-modal rather than paradigm-internal validity evidence.

5.2. Pathway 2: Human-Agent Interpretive Division of Labour

The second pathway concerns the conditions under which human interpretive agency should be retained, augmented, or delegated in agent-assisted research. The evidence reviewed in Section 4.2 suggests that the current norm of full human agency in interpretation is neither necessary nor always desirable, but it does not suggest that human agency is dispensable. The DeTAILS study [18] found that quantitative measures of agreement between AI-generated and human-generated themes did not fully capture interpretive adequacy: researchers reported that AI outputs were sometimes less nuanced, missed culturally embedded references, and occasionally flattened distinctions that they considered analytically important.

Research in this pathway should map the specific conditions under which AI-generated interpretation is likely to be adequate and the conditions under which human interpretive revision is methodologically necessary. This mapping should be empirical rather than principled: it cannot be derived from philosophical commitments about the nature of interpretation, because the empirical record has already shown that those commitments do not reliably predict where AI interpretation succeeds or fails.

5.3. Pathway 3: Reporting Standards for Recursive Research Processes

The third pathway concerns the practical challenge of documenting and communicating research conducted through a recursive, non-linear agent loop. Current reporting standards assume that the researcher can provide a coherent account of a sequence of decisions made in a defined order. Agent-loop research produces a trajectory of decisions that may include automated revisions, abandoned hypotheses, and dynamic data collection choices not planned in advance. Without appropriate reporting standards, this trajectory cannot be evaluated by readers or reviewers.

Code Listing 3 illustrates a proposed provenance record schema that captures the key properties a reporting standard for agent-loop research would need to document. The schema draws directly on the state and transition fields of the queryLoop() architecture described in Section 4.4.

Code Listing 3. Proposed provenance record schema for agent-loop research, illustrating Pathway 3 reporting requirements. Fields capture trajectory, human intervention points, and review procedures.

```
// Code Listing 3 -- Proposed provenance record schema (Pathway 3)
{
  "study_id":                  "AUA-UM-2025-001",
  "agent_architecture":        "PRAO / ReAct",
  "foundation_model":          "claude-sonnet-4-6",
  "stopping_criterion":        "max_iterations=15 OR researcher_approval=true",
  "total_iterations_executed": 9,
  "human_intervention_points": [3, 7],
  "interventions": {
    "3": "Researcher revised sampling criterion for document retrieval",
    "7": "Researcher rejected AI-generated causal interpretation, substituted
own"
  },
  "autonomous_actions": [
    "statistical_model_selection",
    "thematic_codebook_generation",
    "outlier_flagging"
  ],
  "trajectory_hash":           "sha256:4f8b2a...",
  "post_loop_review":          "Full trajectory reviewed by two independent
researchers"
}
```

5.4. Pathway 4: Ethical Governance of Autonomous Methodological Agents

The fourth pathway concerns the ethical implications of delegating methodological decisions to autonomous agents. Research ethics frameworks have developed around the assumption that a human researcher bears responsibility for all decisions made during a study. When an agent autonomously selects a statistical model, identifies and excludes outliers, or generates a thematic interpretation that shapes the study's conclusions, the attribution of methodological responsibility becomes unclear. This is not a hypothetical concern: the agent-loop research pipelines documented in our corpus include systems that made all four of these decisions autonomously in documented studies.

Research in this pathway should examine how existing ethical frameworks, including Institutional Review Board

requirements, research integrity guidelines, and data protection regulations, apply to agent-loop research, and where they require extension or revision. Of particular importance is the question of whether the opacity of LLM reasoning constitutes a form of methodological undisclosure incompatible with current transparency norms. The provenance record proposed in Pathway 3 would partially address this concern, but it would not resolve the deeper question of whether a research community is willing to accept outputs whose interpretive trajectory cannot be fully reconstructed.

6. Conclusion

A problematization review of 94 studies published between 2018 and 2025 shows that three foundational assumptions of the qualitative-quantitative divide are empirically contested by AI agent loop architectures already operating in research settings. The claim that data type determines method selection, that interpretation requires human agency, and that research inference proceeds sequentially are each weakened, in measurable ways, by what production-grade agent loops do in practice. Examination of the Anthropic Claude Code queryLoop() architecture grounds this argument at the level of executable code: its condition-based termination logic, persistent conversation history across the state.messages field, and machine-readable continuation rationale in state.transition correspond directly to the adaptive properties of skilled human interviewing, formalised in a type system that enforces compliance at compile time.

The cross-domain case of speaker diarization optimisation for Sarawak Malay, examined in Section 4.5, extends these findings beyond language model-specific systems. A recursive pseudo-label transfer learning pipeline spanning raw audio waveforms, categorical RTTM annotations, and Diarization Error Rate metrics contests all three assumptions through the same PRAO cycle identified in the text-centric analyses, confirming that the architectural contestation is domain-general rather than a property of any particular modality or model family. Taken together, these findings do not render the qualitative-quantitative divide obsolete. Researchers still face substantive choices about interpretive rigour and causal warrant; what agent loops remove is the technical necessity of treating those choices as binary opposites requiring a paradigmatic declaration at the outset of a study.

The four research pathways proposed in Section 5, covering validity theory for data-type-agnostic inference, human-agent interpretive division of labour, reporting standards for recursive research processes, and ethical governance of autonomous methodological agents, are designed to give researchers conceptual and procedural scaffolding that the current literature does not yet supply. Developing those frameworks requires the kind of cross-paradigm collaboration that the qualitative-quantitative divide has historically impeded. Future work should subject each pathway to empirical testing as agent-loop research deployments become more widely documented in peer-reviewed venues.

AI Usage Statement / Declaration. In the preparation of this manuscript, the authors used large language model tools strictly to assist with language clarity and grammar checking. All analytical content, argumentation, and conclusions are the exclusive intellectual product of the human authors. All AI usage is disclosed in accordance with journal policy.

References

[1] Creswell, J. W. and Creswell, J. D., *"Research Design: Qualitative, Quantitative, and Mixed Methods Approaches, 6th edition,"* SAGE Publications, 2023.

[2] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y, "ReAct: Synergizing Reasoning and Acting in Language Models," *In the Proceedings of International Conference on Learning Representations*, 2023.

[3] Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., and Wen, J. R., "A Survey on Large Language Model Based Autonomous Agents," *Frontiers of Computer Science* 18 (6), 186345 (2024) https://doi.org/10.1007/s11704-024-40231-1

[4] Alvesson, M. and Sandberg, J., "Generating Research Questions Through Problematization," *Academy of Management Review* 36 (2), 247–271 (2011). https://doi.org/10.5465/amr.2009.0188

[5] Lin, W. L. (2026, March 30). How Claude Code Works. https://wanlanglin.github.io/-awesome-cc-harness/ Retrieved 22 April, 2026.

[6] Berger, P., and Luckmann, T., *"The social construction of reality,"* in Social Theory Re-Wired, pp. 110-122, Routledge, 2016.

[7] Orlikowski, W. J. and Baroudi, J. J., "Studying Information Technology in Organizations: Research Approaches and Assumptions," *Information Systems Research* 2 (1), 1–28 (1991). https://doi.org/10.1287/isre.2.1.1

[8] Creswell, J. W., and Clark, V. L. P., *"Designing and Conducting Mixed Methods Research,"* Sage Publications, 2017.

[9] Morgan, D. L., *"Integrating Qualitative and Quantitative Methods: A Pragmatic Approach,"* Sage Publications, 2013.

[10] Mahdi, H., "Perceive, Plan, Act, Self-Correct: An Architectural Framework for Goal-Directed Agentic AI Systems," engrXiv (2025). https://doi.org/10.31224/6738

[11] Wu, Q., Bansal, G., Zhang, J., Wu, Y., Zhang, S., Zhu, E., Li, B., Jiang, L., Zhang, X., and Wang, C., "AutoGen: Enabling next-gen LLM applications via multi-agent conversation," arXiv preprint arXiv:2308.08155 (2023). https://doi.org/10.48550/arXiv.2308.08155

[12] LangChain Python Overview, https://docs.langchain.com/oss/python/langchain/overview Retrieved 22 April, 2026.

[13] Abou Ali, M., Dornaika, F., and Charafeddine, J., "Agentic AI: A Comprehensive Survey of Architectures, Applications, and Future Directions," *Artificial Intelligence Review* 59 (1), 11 (2025). https://doi.org/10.1007/s10462-025-11422-4

[14] Anthropic. (2025). What 81,000 people want from AI. Anthropic. https://www.anthropic.com/features/81k-interviews. Retrieved 22 April, 2026.

[15] Braun, V. and Clarke, V., "Reflecting on reflexive thematic analysis," *Qualitative Research in Sport, Exercise and Health* 11 (4), 589–597 (2019). https://doi.org/10.1080/2159676X.2019.1628806

[16] Morgan, D. L., "Exploring the Use of Artificial Intelligence for Qualitative Data Analysis: The Case of ChatGPT," *International Journal of Qualitative Methods* 22, 16094069231211248 (2023). https://doi.org/10.1177/16094069231211248

[17] Venkatesh, V., Brown, S. A., and Sullivan, Y. W., "Guidelines for Conducting Mixed-Methods Research: An Extension and Illustration," *Journal of the Association for Information Systems* 17 (7), 2 (2016). DOI: 10.17705/1jais.00433

[18] Sharma, A. and Wallace, J. R., "DeTAILS: Deep Thematic Analysis with Iterative LLM Support," *In the Proceedings of the 7th ACM Conference on Conversational User Interfaces*, ACM Press, 1–7 (2025). https://doi.org/10.1145/3719160.3735657

[19] Karjus, A., "Machine-Assisted Quantitizing Designs: Augmenting Humanities and Social Sciences with Artificial Intelligence," *Humanities and Social Sciences Communications* 12 (1), 1–18 (2025). https://doi.org/10.1057/s41599-025-04503-w

[20] Zhang, H., Wu, C., Xie, J., Lyu, Y., Cai, J., and Carroll, J. M., "Redefining Qualitative Analysis in the AI Era:

Utilizing ChatGPT for Efficient Thematic Analysis," arXiv Preprint (2023). https://doi.org/10.1016/j.chbah.2025.100144

[21] Klein, H. K., and Myers, M. D., "A Set of Principles for Conducting and Evaluating Interpretive Field Studies in Information Systems," *MIS Quarterly* 23 (1), 67–93 (1999). https://doi.org/10.2307/249410

[22] Hitch, D. (2024). "Artificial intelligence augmented qualitative analysis: the way of the future?," *Qualitative Health Research* 34 (7), 595–606. https://doi.org/10.1177/10497323231217392

[23] Bennis, I., Mouwafaq, S., "Advancing AI-driven thematic analysis in qualitative research: a comparative study of nine generative models on Cutaneous Leishmaniasis data," *BMC Medical Informatics and Decision Making* 25 (1), 124 (2025). https://doi.org/10.1186/s12911-025-02961-5

[24] Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., and Ha, D., "The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery," arXiv preprint arXiv:2408.06292 (2024). https://doi.org/10.48550/arXiv.2408.06292

[25] Ghafarollahi, A., and Buehler, M. J., "SciAgents: Automating Scientific Discovery through Bioinspired Multi-Agent Intelligent Graph Reasoning," *Advanced Materials* 37 (22), 2413523 (2025). https://doi.org/10.1002/adma.202413523

[26] Luo, Z., Yang, Z., Xu, Z., Yang, W., and Du, X., "LLM4SR: A Survey on Large Language Models for Scientific Research," arXiv preprint arXiv:2501.04306 (2025). https://doi.org/10.48550/arXiv.2501.04306

[27] Liu, M., Liang, K., Zhao, Y., Tu, W., Zhou, S., Gan, X., and He, K., "Self-Supervised Temporal Graph Learning with Temporal and Structural Intensity Alignment," *In IEEE Transactions on Neural Networks and Learning Systems,* IEEE Press, 6355–6367 (2024). https://doi.org/10.1109/TNNLS.2024.3386168

[28] Timmermans, S., and Tavory, I., "Theory Construction in Qualitative Research: From Grounded Theory to Abductive Analysis," *Sociological Theory* 30 (3), 167–186 (2012). https://doi.org/10.1177/0735275112457914

[29] Chen, H., Lei, Y., Zhang, D., Ke, B., Zhu, D., Chen, X., and Wang, H., "MatryoshkaThinking: Recursive Test-Time Scaling Enables Efficient Reasoning," arXiv preprint arXiv:2510.10293 (2025). https://doi.org/10.48550/arXiv.2510.10293

[30] Gao, J., Guo, Y., Lim, G., Zhang, T., Zhang, Z., Li, T. J. J., and Perrault, S. T., "CollabCoder: A Lower-Barrier, Rigorous Workflow for Inductive Collaborative Qualitative Analysis with Large Language Models," In the Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, ACM Press, 1–29 (2024). https://doi.org/10.1145/3613904.3642002

[31] Rahim, M. Z., Juan, S. S., and Mohamad, F. S., "Improving Speaker Diarization for Low-Resourced Sarawak Malay Language Conversational Speech Corpus," In the Proceedings of 2023 International Conference on Asian Language Processing (IALP), IEEE Press, 228–233 (2023). https://doi.org/10.1109/IALP61005.2023.10337314