Association for Information Systems

AIS Electronic Library (AISeL)

ICIS 2026 Proceedings

International Conference on Information Systems (ICIS)

December 2026

The Methodological Black Box of LLM-Assisted Topic Modeling

Carlotta Cochis
University of Modena and Reggio Emilia, carlotta.cochis@unimore.it

Follow this and additional works at: https://aisel.aisnet.org/icis2026

Recommended Citation
Cochis, Carlotta, "The Methodological Black Box of LLM-Assisted Topic Modeling" (2026). *ICIS 2026 Proceedings*. 15.
https://aisel.aisnet.org/icis2026/res_methods/res_methods/15

This material is brought to you by the International Conference on Information Systems (ICIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in ICIS 2026 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

The Methodological Black Box of LLM-Assisted Topic Modeling

Short Paper

Carlotta Cochis
Department of Sciences and Methods for Engineering
University of Modena and Reggio Emilia
Reggio Emilia, Italy
carlotta.cochis@unimore.it

Abstract

Generative AI is increasingly delegated analytical tasks in research workflows, yet what large language models actually do when asked to "run topic modeling" remains opaque. We compare five consumer LLM systems (ChatGPT, Claude, DeepSeek, Gemini, Grok) performing topic modeling on the same corpus of 77 articles through one-shot, deliberately underspecified delegation, in November 2025 and again — three same-instruction runs and two paraphrases per system — in August 2026. Systems diverged at the level of the algorithm family, not merely its parameters; divergence recurred across runs, versions, and phrasings, and included form-preserving methodological substitution, undisclosed alternatives, and simulated results acknowledged only under disclosure probing. We propose the Six-Gate Framework for Accountable LLM-Assisted Topic Modeling. The framework decomposes the task into six methodological decision gates and replaces monolithic delegation with adaptive human–AI dialogue and explicit researcher authorization. Its ongoing validation compares the autonomous baseline with a framework-guided condition.

Keywords: topic modeling; large language models; research methods; human–AI collaboration; methodological transparency; systematic literature reviews

Introduction

Generative AI tools are being rapidly adopted in management and information systems research, including for systematic literature reviews. Among their applications, topic modeling—the discovery of latent thematic structures in research corpora—is one of the most prominent (Hannigan et al., 2019; Khraisha et al., 2024). The promise — natural-language access to methods that previously required programming and statistical training — is substantial; the risks remain underexamined.

Topic modeling is methodologically demanding: the researcher must select an algorithm, tune hyperparameters such as the number of topics (k), validate outputs interpretively, and document each choice. When these decisions are delegated to an LLM without explicit constraint, what does the LLM actually do? The few studies that have begun to examine AI-assisted research synthesis have focused on output quality (Khraisha et al., 2024) but have not systematically characterized the methodological choices that LLMs make autonomously, nor the consistency of those choices across models. Our study addresses this gap. We focus on two research questions. RQ1: Under one-shot, deliberately underspecified delegation of a topic modeling task, what methodological choices do contemporary LLMs make, and how consistent are those choices across models and runs? RQ2: Does a framework-guided interaction that decomposes the task into six decision gates and preserves adaptive human–AI dialogue improve convergence, validity, and auditability relative to autonomous one-shot delegation?

To address these questions, we compare two interaction architectures using the same five state-of-practice LLMs and the same corpus of 77 peer-reviewed articles on data governance and AI. In the autonomous baseline condition, each model receives the complete analytical task through a single deliberately underspecified instruction and resolves the methodological choices independently. In the framework-

guided condition, the same task is decomposed across the Six-Gate Framework: the methodological objective and the researcher's decision authority are fixed at each gate, while the dialogue adapts to each system's responses. The completed baseline establishes the comparison against which the framework is evaluated. Using the same corpus and an intentionally vague instruction, five large language models employed four distinct methodological families. One model preserved the appearance of topic modeling while replacing the methodology with deterministic keyword matching against predefined categories—what we term form-preserving methodological substitution. The variation in the number of topics (from 4 to 25) highlights a deeper issue: task-level epistemic divergence, where identical task labels lead to fundamentally different methods across models.

This paper makes three contributions. First, it documents empirically the methodological divergence that emerges under autonomous, one-shot delegation, extending prior accounts of AI variability beyond parameter dispersion to fundamental methodological commitment. Second, it develops the Six-Gate Framework for Accountable LLM-Assisted Topic Modeling as an interaction architecture that decomposes the analytical task and requires explicit researcher authorization at each consequential decision point. The framework standardizes decision obligations rather than prompts. Third, it specifies an empirical evaluation through a within-system comparison between the autonomous baseline and the framework-guided condition, testing effects on convergence, validity, auditability, and replicability.

Theoretical Background

Topic Modeling in Management and IS Research

Topic modeling is an unsupervised machine learning approach for discovering latent thematic structures in unstructured text, established as a methodological standard for systematic literature reviews in management and information systems research (Hannigan et al., 2019). Within IS research specifically, topic modeling has been used to map the discipline's intellectual core (Sidorova et al., 2008), codified in dedicated tutorials for IS researchers (Debortoli et al., 2016), and methodologically extended in the field's leading journals (Yang et al., 2023), while the automation of literature reviews — the delegation scenario studied here — is an active IS research stream in its own right (Wagner et al., 2022). Latent Dirichlet Allocation (LDA), the dominant probabilistic algorithm, models documents as mixtures of latent topics and topics as distributions over words (Blei et al., 2003). Implementation requires the researcher to specify the number of topics (k) using statistical criteria such as perplexity, harmonic mean, or coherence scores (Griffiths & Steyvers, 2004), to select preprocessing parameters, and to validate output through interpretive judgment.

Generative AI as an Analytical Agent

Large language models operate fundamentally differently from purpose-built statistical procedures. As autoregressive systems trained to predict the most probable next token given prior context (Brown et al., 2020), LLMs do not perform statistical computation in the classical sense unless they invoke external tools (e.g., code execution environments). When prompted to perform an analytical task, an LLM may either (i) generate executable code that implements a statistical procedure, (ii) orchestrate an external pipeline through tool calls, or (iii) produce output that mimics the structure of an analytical result without implementing any underlying procedure. The methodological consequences of these three modes differ substantially, but the distinction is not visible from the output alone—a problem we revisit empirically. Emerging evidence indicates that LLM outputs exhibit known limitations in long-context processing, statistical reasoning, and calibration of uncertainty (Dell'Acqua et al., 2026). When LLMs nonetheless produce outputs that resemble valid analyses, the resulting risk is not error per se but *invisible methodological commitment*: the model selects an algorithm, a parameter, a preprocessing pipeline, without making these choices legible to the researcher.

Human–AI Collaboration in Knowledge Work

The integration of LLMs into research workflows has prompted a growing body of work on human–AI collaboration in knowledge-intensive tasks (Seeber et al., 2020; Bansal et al., 2021; Dell'Acqua et al., 2026). Two findings from this literature are particularly relevant. First, complementary performance—where the human–AI team outperforms either alone—is not automatic; it requires structured division of labor that aligns human and AI strengths and exposes the AI's weaknesses to human oversight (Bansal et al., 2021).

Second, evidence from large-scale field experiments indicates that human oversight is more effective when it intervenes at decision points internal to the analytical workflow than when it is delayed until after the AI has produced an output (Dell'Acqua et al., 2026). This second finding directly motivates the framework's six decision points. How to structure the division of labor between researcher and LLM in tasks requiring sustained methodological judgment remains an open question (Khraisha et al., 2024).

A Six-Gate Framework for Accountable LLM-Assisted Topic Modeling

The autonomous baseline findings (next section) reveal that LLMs cannot be treated as neutral executors of a well-defined task: they make consequential methodological choices silently, and these choices diverge systematically across models. We propose a six-gate framework (Figure 1) that responds to this finding by structuring the human–AI interaction around disambiguation, justification, and auditability at each consequential decision point. No analytical commitment passes a gate until the researcher has reviewed the surfaced alternatives and explicitly authorized one. The gates standardize what must be surfaced, decided, and recorded; they do not prescribe a universal sequence of prompts. Dialogue within each gate remains adaptive to the LLM's responses, omissions, questions, and evidence, while the researcher retains binding decision authority. The framework departs from prior workflow proposals (Khraisha et al., 2024; Hannigan et al., 2019) by elevating task definition itself to the first decision point, rather than treating it as implicit common ground between researcher and tool.

Stage 1: Task Definition

Gate 1 — Task Disambiguation. Before any analysis is initiated, the researcher must explicitly specify the intended methodological family. The AI's role at this gate is to surface the available methodological families, articulate the trade-offs, and flag any ambiguity in the researcher's request. The researcher decides which family is appropriate to the research question. This gate prevents the silent commitment to incommensurable methods documented in the autonomous baseline.

Gate 2 — Corpus Assessment. The AI conducts exploratory analysis (vocabulary size, document length distribution, temporal coverage, language consistency) and reports indicators relevant to the chosen methodological family. The researcher validates corpus suitability and identifies any preprocessing requirements that the chosen family imposes. This gate prevents "garbage-in–garbage-out" failures.

Stage 2: Methodological Choices

Gate 3 — Algorithm Specification. Within the family selected at Gate 1, the researcher and AI jointly specify the concrete algorithm (e.g., LDA versus CTM within probabilistic generative modeling). The AI is asked to surface alternatives it would otherwise leave implicit—addressing directly the "hidden alternatives" pattern observed in the autonomous baseline—and to provide explicit justification for each candidate. The researcher selects.

Gate 4 — Parameter Determination. The AI reports candidate values for the key hyperparameters (k, alpha, beta in the LDA case) using multiple metrics (e.g., perplexity, harmonic mean, coherence scores, silhouette), explicitly noting where metrics disagree. The researcher decides the final values, motivating any deviation from metric-suggested optima with reference to interpretability or theoretical fit.

Stage 3: Execution and Audit

Gate 5 — Execution and Validation. The AI executes the specified algorithm with the specified parameters. The researcher then evaluates the output along established dimensions — interpretability, coverage, distinctiveness, keyword quality, assignment confidence (Hannigan et al., 2019). Based on this evaluation, the researcher decides whether to accept the results, iterate (returning to Gates 3 or 4), or reject the approach and reconsider Gate 1.

Gate 6 — Auditability. The AI documents all methodological decisions made throughout the workflow: rationales, algorithm and parameter choices, metrics consulted, human overrides, and iterations. The researcher reviews and signs off on the audit trail. This documentation accompanies any resulting publication, enabling reviewers to inspect not only what was done but on what basis.

The framework clearly allocates responsibility: *the AI executes and recommends; the human decides and accounts for*. This allocation locates accountability with the human user, consistent with emerging

regulatory principles for AI-assisted decision-making (European Union, 2024), and provides the explicit human warrant that AI-assisted research otherwise lacks.

Gate	AI task (recommends/executes)	Human decision (accountable)
STAGE 1: Task Definition		
Gate 1: Task Disambiguation	AI: Surface methodological families and trade-offs	Human: Selects the methodological family
Gate 2: Corpus Assessment	AI: Conduct exploratory analysis of corpus	Human: Validates corpus suitability
STAGE 2: Methodological Choices		
Gate 3: Algorithm Specification	AI: Surface algorithmic alternatives with justifications	Human: Selects the concrete algorithm
Gate 4: Parameter Determination	AI: Compute candidate values via multiple metrics	Human: Sets final parameter values
STAGE 3: Execution and Audit		
Gate 5: Execution and Validation	AI: Execute algorithm; report output	Human: Evaluates and decides accept/iterate/reject
Gate 6: Auditability	AI: Document all decisions and rationales	Human: Reviews and signs off audit trail

Figure 1. The Six-Gate Framework for Accountable LLM-Assisted Topic Modeling in Systematic Literature Reviews.

Empirical Evidence: Autonomous Baseline Condition

Research Design

To examine whether large language models operate as reliable autonomous agents for topic modeling in systematic literature reviews, we conducted a comparative analysis under conditions of full methodological autonomy. Five LLMs—accessed through their publicly available web interfaces between November 7 and November 11, 2025—received the same deliberately underspecified instruction to perform topic modeling on a research corpus, with no constraints imposed on algorithm choice, parameter selection, preprocessing decisions, or output structure.

The five LLMs analyzed were ChatGPT (GPT-5), Claude (Sonnet 4.5), DeepSeek (V3.1), Gemini (2.5 Pro), and Grok (Grok 4). These five systems were selected as the most widely used consumer chatbots at the time of data collection, accessed through their public web interfaces rather than APIs: the design targets the delegation scenario as researchers actually encounter it rather than a controlled capability benchmark, and heterogeneity in release dates and capability tiers is therefore part of the condition under study rather than a confound. An experienced researcher developed a human expert baseline in parallel using established R-based topic modeling procedures. All models received identical input: an Excel file containing 77 peer-reviewed articles on the intersection of data governance and artificial intelligence, retrieved from Scopus using the query (TITLE-ABS-KEY("data governance" OR "data policy" OR "data ecosystem" OR "information system governance" OR "information governance") AND TITLE-ABS-KEY("AI" OR "artificial intelligence" OR "ML" OR "machine learning")) with filters limiting results to business subject area, English language, and article/review/editorial document types. The query returned 83 records: five were excluded at relevance screening because they did not address data governance and AI jointly, and one further record — a non-empirical practitioner briefing — was removed at final curation, yielding the final corpus of 77 articles. The expert baseline modeled, for each of the 77 records, the concatenation of Title, Abstract, and Author and Index Keywords (median 235 words per record, min 10, max 582). The prompt deliberately left the choice of text fields to each model — a further silent degree of freedom, which models resolved differently, as surfaced in the disclosure sequences. Baseline preprocessing followed the standard tm pipeline: lowercasing, removal of numbers and punctuation, English and custom stop-word lists, a minimum token length of three characters, sparsity-based term pruning, and no stemming or lemmatization.

The expert baseline was specified ex ante (November 2025) by triangulating selection criteria over k∈[2,50]: the harmonic-mean estimator peaked at k=22, held-out perplexity (80/20) reached its minimum near k=36, and the Arun et al. (2010) diagnostic favored far smaller solutions (global minimum k=2, local minimum k=13). As these criteria genuinely disagree, we selected the likelihood optimum (k=22) as the granularity most suitable for downstream comparability, after interpretive inspection of candidate solutions (k=13, 15, 22, 36), and refit the model under five random seeds, retaining the median-perplexity run. Coherence-based selection on the same corpus (C_v and NPMI, top-10 terms) favors far coarser solutions (optima at k=2, NPMI ≈ 0 beyond k≈10), aligning with the Arun diagnostic. We therefore treat the expert model as one defensible instantiation of the canonical pipeline — a reference point rather than ground truth — and the framework-guided comparison will report agreement metrics against both this anchor and a coherence-favored coarse anchor as a sensitivity analysis.

The prompt (Figure 2) did not specify any algorithm, parameter, evaluation metric, or preprocessing approach—these decisions were left entirely to each model. This underspecification is deliberate: it defines the autonomous baseline condition and exposes the default behavior each system enacts when the method is left unstated. The study's intervention is not a more detailed one-shot prompt, but framework-guided task decomposition and adaptive dialogue evaluated against this baseline. To verify what each model had actually executed, follow-up prompts requested disclosure of the analytical environment, code, and methodological rationale used. We archived all conversations and outputs for reproducibility. To separate stable dispositions from sampling variation, we re-executed the protocol in August 2026 on the same five systems in their then-current consumer versions (one newly created account per system, chat memory disabled, default reasoning settings). Each system received the identical frozen prompt in three independent sessions (R1–R3) and two paraphrases in further sessions: P1 rephrases the request while retaining the term "topic modeling"; P2 asks for latent-theme identification without naming the technique and renames the output schema accordingly. All twenty delivered workbooks passed structural validation (77/77 documents, no duplicates, no fabricated identifiers), and every session ended with the same four-step disclosure sequence (provenance, method and parameters, alternatives considered, assumptions; Figure 2). Because several systems were updated between November and August, cross-wave comparisons are product-level rather than version-level; the within-wave repetitions isolate run-to-run variability at a fixed point in time.

(a) Task instruction (November 2025; frozen for the August 2026 runs R1–R3)
The attached dataset contains 77 papers. Run topic modeling on these papers. The dataset includes columns for Authors, Title, Year, Abstract, Author Keywords, DOI, and EID. Determine the optimal number of topics and provide results in an Excel file with two sheets: (1) topics_[model] containing model, topic_id, label, k1-k10 keywords, and prevalence; (2) assign_[model] containing model, eid, doi, assigned_topic_id, assigned_topic_label, topic_score, and notes.
(b) Paraphrases (August 2026; one session each)
P1 — term retained. *The attached dataset contains 77 papers. The dataset includes columns for Authors, Title, Year, Abstract, Author Keywords, DOI, and EID. Perform a topic modeling analysis of these papers, choosing the number of topics you consider optimal, and provide results in an Excel file with two sheets: (1) topics_[model] containing model, topic_id, label, k1-k10 keywords, and prevalence; (2) assign_[model] containing model, eid, doi, assigned_topic_id, assigned_topic_label, topic_score, and notes.* **P2 — term avoided.** *The attached dataset contains 77 papers. The dataset includes columns for Authors, Title, Year, Abstract, Author Keywords, DOI, and EID. Identify the main latent themes in these papers and classify each paper under the theme that best describes it, deciding yourself how many themes are appropriate. Provide results in an Excel file with two sheets: (1) themes_[model] containing model, theme_id, label, k1-k10 keywords, and prevalence; (2) assign_[model] containing model, eid, doi, assigned_theme_id, assigned_theme_label, theme_score, and notes.*
(c) Post-delivery disclosure sequence (August 2026; identical in every session, sent one message at a time)
D1 — provenance. *Disclose exactly how you produced this analysis. (a) Did you execute code to produce these results? (b) If yes, provide the exact code you executed, unmodified - not a cleaned-up or reconstructed version. If you cannot retrieve the exact code you executed, state this explicitly. (c) Describe the execution environment: language, libraries, versions. (d) If you did not execute code, describe precisely, step by step, the procedure you actually followed.*

D2 — method and parameters. *State the specific analytical method you used (e.g., LDA, NMF, embedding-based clustering, graph community detection, keyword-based classification, or other). Report all parameter values, including the number of topics and the criterion used to select it, and list all preprocessing steps applied.*
D3 — alternatives considered. *Before committing to this approach, did you consider alternative methods or configurations? List every alternative you considered and explain why you rejected each one. If you considered none, state so explicitly.*
D4 — assumptions. *What ambiguities, if any, did you perceive in my original request, and what assumptions did you make to resolve them?*

Figure 2. Verbatim task instructions and post-delivery disclosure sequence (autonomous baseline condition).

Note: These sessions apply no framework gate and constitute the autonomous baseline; the framework-guided condition is now in execution. The August 2026 runs re-used the November instruction verbatim. The disclosure sequence follows the deliverable and therefore documents, without influencing, the session's methodological choices; when a promised file was not delivered, one scripted retrieval probe followed. Any clarification question from a model received the same scripted reply: "Proceed as you consider appropriate."

Results: Methodological Divergence Across Models

Given the same corpus and the same deliberately underspecified instruction, the five LLMs executed four fundamentally different methodological approaches. Only two of five models implemented Latent Dirichlet Allocation, the canonical probabilistic topic modeling algorithm employed by the human expert. One model substituted Non-negative Matrix Factorization. One model implemented graph-based community detection on a keyword co-occurrence network. One model performed deterministic supervised classification using LLM-defined topic categories—a procedure that does not constitute topic modeling in any standard methodological sense. Table 1 summarizes the algorithm family and topic granularity delivered by every run in both waves.

Agent	Nov 2025	Aug R1	Aug R2	Aug R3	Aug P1	Aug P2
Human expert	LDA k=22	n/a	n/a	n/a	n/a	n/a
Claude	LDA k=8	NMF k=6	NMF k=6	NMF k=6	NMF k=6	NMF k=7
DeepSeek	LDA k=5	—	—	—	—	—
ChatGPT	NMF k=9	NMF k=5	NMF k=12	NMF k=8	NMF k=7	NMF k=9
Gemini	KM k=4	LDA k=5	LDA k=3	LDA k=5	LDA k=5	LDA k=5
Grok	CD k=25	NMF k=6	LDA k=9	LDA k=6	LDA k=7	HYB k=6

Table 1. Methodological Choices Across LLMs in the Autonomous Baseline Condition (November 2025 submitted run and August 2026 replications; n=77 documents).

Note. Cells report the delivered analysis: algorithm family and number of topics (LDA = latent Dirichlet allocation; NMF = non-negative matrix factorization; KM = LLM-defined keyword matching; CD = graph community detection; HYB = rule-based/TF-IDF hybrid classification). November 2025: single submitted run. August 2026 (16–23 August): current consumer versions, default settings, one dedicated account per system; R1–R3 repeat the identical frozen prompt in independent sessions, P1 paraphrases it retaining the term "topic modeling," P2 requests latent themes without naming the technique. — = no analysis file delivered (simulation acknowledged under disclosure); n/a = the expert baseline is a single ex-ante instantiation, not re-run.

The dispersion is not limited to the parameter k, which ranges from 4 to 25 across LLMs in the November 2025 run (coefficient of variation 67.7%). The dispersion extends to the choice of method family, the criterion for parameter selection, the vocabulary of evaluation metrics, the preprocessing pipeline (each LLM defined its own custom domain stopword list), and the output completeness (Grok left 10.4% of papers unassigned).

The August 2026 replications separate two layers of stability. Within a fixed version and identical instructions (R1–R3), the algorithm family was stable for Claude (NMF), Gemini (LDA), and ChatGPT (NMF), but not for Grok (NMF in R1, LDA thereafter); DeepSeek delivered no analysis file in any of the five sessions and, under disclosure probing, acknowledged that displayed results were simulated rather than computed (e.g., A2: *"I must be transparent with you… manually processed the data in my 'mind'…"* · A3: *"NO. I did not execute any code…")*. Topic granularity, by contrast, varied within model even under identical instructions — ChatGPT selected k=5, 12, and 8 across its three runs — while Claude held k=6 in every same-instruction run. Paraphrases largely preserved each model's within-wave family, with one exception:

when the request avoided the term "topic modeling" (P2), Grok substituted a rule-based/TF-IDF hybrid classification — the same form-preserving substitution pattern observed for Gemini in November. Across waves, ChatGPT was the only system that retained its November family (NMF): Claude moved from LDA to NMF, DeepSeek from an executed gensim LDA to non-execution, Gemini from keyword matching to LDA, and Grok from community detection to NMF/LDA. Run-to-run divergence and version-to-version drift are therefore distinct phenomena: the former bounds what any single session can establish; the latter is the temporal non-stationarity that motivates a version-matched comparison between the autonomous and framework-guided conditions.

The Hidden Alternatives Problem

A second finding emerged from the follow-up disclosure prompts and warrants explicit attention. When asked to disclose its analytical reasoning, Claude reported considering TF-IDF with K-means clustering as an alternative to LDA before ultimately defaulting to the latter. The model described TF-IDF as potentially better suited to a corpus of the size and lexical density of the one analyzed, yet did not surface this consideration in its initial output. The researcher received a result accompanied by no indication that alternatives had been internally entertained, evaluated, and rejected. We label this pattern *the hidden alternatives problem*: LLMs may carry methodological reasoning richer than the reasoning they expose by default, and the researcher receives a single recommendation rather than a comparison among the candidates the model itself considered. This pattern matters because it obscures the deliberation that distinguishes a methodological choice from an arbitrary one. Default outputs alone cannot determine whether other LLMs in our sample exhibited similar internal deliberation; this opacity is itself the problem. We therefore treat hidden alternatives as a case-based candidate failure mode rather than an established regularity: retrospective self-report cannot establish its prevalence—disclosure is generative, not archival—while Gate 3 converts it into a directly observable quantity.

Discussion

The autonomous baseline reveals a previously underdocumented failure mode in LLM-assisted research: task-level underspecification. When asked to "run topic modeling," none of the five models converged on what the task means at the methodological level. Three plausible mechanisms account for this divergence.

First, *semantic ambiguity in the task label*. "Topic modeling" is an umbrella term that legitimately spans several unsupervised families — probabilistic generative models (LDA), matrix factorization (NMF), embedding-based clustering, graph-based community detection — and LLMs may additionally misconstrue it as licensing rule-based supervised classification. Each LLM appears to have selected from this interpretive space according to implicit priors learned during training, with no shared methodological canon enforcing convergence. The Gemini case highlights a problematic pattern called form-preserving methodological substitution, where a model produces output that appears similar to standard topic modeling—such as topics with keywords, document-topic assignments, and distribution statistics—but rests on a procedure that does not constitute topic modeling under any standard definition. While both probabilistic modeling and matrix factorization infer latent structures, Gemini's approach enforces a pre-defined structure and categorizes documents through keyword matching, preventing discovery of new themes. The results look correct while concealing the underlying methodological change — a hard-to-detect failure mode precisely because the output owes its apparent authenticity to its correct shape, with the fully specified output format supplying the surface that enables concealment.

Second, *default helpfulness over clarification*. None of the five models requested clarification on what "topic modeling" should mean in this context, despite the obvious ambiguity. This pattern is consistent with documented tendencies of contemporary LLMs to optimize for task completion over inquiry (Dell'Acqua et al., 2026; Bansal et al., 2021). For methodologically sensitive tasks, this default is hazardous: the model commits to an arbitrary methodological position before the researcher is aware that a position needs to be taken. The asymmetry is constitutive of lay delegation: the researcher specifies what they can see — the deliverable and its format — not what they cannot see — the method; and it extends to the input itself, since the prompt lists the available fields without prescribing which ones to model.

Third, *opacity of internal alternatives*. The Claude case—where the model internally considered TF-IDF with K-means clustering before defaulting to LDA, but did not surface this consideration unprompted—suggests that LLMs may carry richer methodological reasoning than they expose in their default outputs.

These mechanisms have direct implications for information systems methodology. First, the principle of methodological transparency—long established in IS research (Webster & Watson, 2002; Hannigan et al., 2019)—requires extension to the task definition itself. It is no longer sufficient to report which algorithm and which parameters were used; researchers must also report which interpretation of the task the AI agent committed to, and on what basis. Second, the assumption that human guidance can be operationalized through parameter specification alone is empirically untenable. Specifying k=22 to a model that implements graph community detection rather than LDA does not yield comparable results across LLMs—it yields incommensurable results. Third, the auditability of AI-assisted research depends on the ability to inspect not only what the AI produced but what the AI did to produce it; this audit trail must include the methodological choices that the AI made silently on the researcher's behalf.

These findings motivate the framework presented in the previous section, and each of the three failure modes maps directly to a specific gate in the framework. *Semantic ambiguity in the task label* is addressed by Gate 1 (Task Disambiguation), which requires the methodological family to be specified before any analysis is initiated, eliminating the latitude that leads to incommensurable executions across LLMs. *Form-preserving methodological substitution*, the most consequential failure mode, is addressed jointly by Gate 1 (which constrains the family of allowable procedures) and Gate 6 (Auditability), which makes the methodological commitment legible in the documentation produced. *The hidden alternatives problem* is addressed in Gate 3 (Algorithm Specification), where the AI is explicitly tasked with surfacing alternatives and providing justification for each, rather than silently committing to a single option. The framework is thus a structured response to these empirically documented failure modes. What is at stake is not error prevention alone: a delegation that passes these gates would turn LLM-assisted topic modeling into a dependable instrument for corpus mapping and gap identification in literature reviews.

Framework Validation: The Framework-Guided Condition

The work now in progress implements and evaluates the framework-guided condition before ICIS 2026. This is not a separate theoretical phase, but the intervention condition in the same comparative design. It tests whether replacing monolithic delegation with six-gate task decomposition and adaptive, researcher-led dialogue reduces the methodological divergence documented in the autonomous baseline while improving validity and auditability. This study compares autonomous and framework-guided conditions across five systems, with three runs per system per condition. The August 2026 replications serve as the autonomous baseline, with interface version strings unchanged during guided-condition execution; otherwise, the autonomous condition is re-executed within the same version window. In the framework-guided setup, the six gates, their informational objectives, and the researcher's binding decision rules are specified ex ante, but dialogue wording, number, and sequence are flexible. Researchers adapt probes and questions based on system responses until gates are satisfied. Comparability relies on common gates and criteria, not identical prompts. All interactions are archived and analyzed to identify patterns and failure modes, which are outcomes rather than preset scripts. Outcomes are measured across four areas: consistency of methodological choices, agreement with an expert or coarse coherence anchor, interpretive validity by blinded experts, and completeness of the audit trail. Some framework-guided runs are repeated on a larger dataset. The design is falsifiable—if guided runs merely mimic the autonomous baseline, the gates would be redundant.

This short paper acknowledges three main limitations. First, the dataset is small and domain-specific, consisting of 77 bibliographic records with a median of 235 words each, a short-text regime in which probabilistic topic models are fragile for any analyst; the divergence observed is at the method-family level and, because all systems received the same corpus, corpus size cannot account for it; whether divergence would attenuate on larger corpora remains untested pending the second-corpus replication. Second, prompt formulation was only varied within a narrow scope—specifically, two paraphrases in the August 2026 replications—and the structured Excel output, part of the delegation contract, may interact with underspecification, making it difficult to disentangle their effects. Third, model versions evolve rapidly, so findings from November 2025 reflect products that have since changed, necessitating cross-wave comparisons at the product level within a single version window to ensure consistency.

AI Usage Statement

The five large language models analyzed in this study (ChatGPT GPT-5, Claude Sonnet 4.5, DeepSeek V3.1, Gemini 2.5 Pro, Grok 4) are the empirical objects of investigation; their use as autonomous topic-modeling agents constitutes the autonomous baseline condition and is documented in detail in the Empirical Evidence section. Beyond this experimental use, generative AI tools assisted with language editing and structural review of intermediate drafts; all methodological choices, theoretical positioning, empirical interpretations, and claims of contribution are the sole responsibility of the author. No AI tool is listed as an author, in compliance with the ICIS 2026 AI Usage Policy.

References

Arun, R., Suresh, V., Veni Madhavan, C. E., & Narasimha Murthy, M. N. (2010). On finding the natural number of topics with latent Dirichlet allocation: Some observations. In Advances in Knowledge Discovery and Data Mining (PAKDD 2010), Lecture Notes in Computer Science, Vol. 6118 (pp. 391–402). Springer. https://doi.org/10.1007/978-3-642-13657-3_43

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Article 81, pp. 1–16). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445717

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

Debortoli, S., Müller, O., Junglas, I., & vom Brocke, J. (2016). Text mining for information systems researchers: An annotated topic modeling tutorial. Communications of the Association for Information Systems, 39, Article 7. https://doi.org/10.17705/1CAIS.03907

Dell'Acqua, F., McFowland, E., III, Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K. C., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2026). Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. *Organization Science, 37*(2), 403–423. https://doi.org/10.1287/orsc.2025.21838

European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L, 2024/1689. http://data.europa.eu/eli/reg/2024/1689/oj

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*(suppl_1), 5228–5235. https://doi.org/10.1073/pnas.0307752101

Hannigan, T. R., Haans, R. F. J., Vakili, K., Tchalian, H., Glaser, V. L., Wang, M. S., Kaplan, S., & Jennings, P. D. (2019). Topic modeling in management research: Rendering new theory from textual data. *Academy of Management Annals, 13*(2), 586–632. https://doi.org/10.5465/annals.2017.0099

Khraisha, Q., Put, S., Kappenberg, J., Warraitch, A., & Hadfield, K. (2024). Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *Research Synthesis Methods, 15*(4), 616–626. https://doi.org/10.1002/jrsm.1715

Seeber, I., Bittner, E., Briggs, R. O., de Vreede, T., de Vreede, G.-J., Elkins, A., Maier, R., Merz, A. B., Oeste-Reiß, S., Randrup, N., Schwabe, G., & Söllner, M. (2020). Machines as teammates: A research agenda on AI in team collaboration. *Information & Management, 57*(2), Article 103174. https://doi.org/10.1016/j.im.2019.103174

Sidorova, A., Evangelopoulos, N., Valacich, J. S., & Ramakrishnan, T. (2008). Uncovering the intellectual core of the information systems discipline. MIS Quarterly, 32(3), 467–482. https://doi.org/10.2307/25148852

Wagner, G., Lukyanenko, R., & Paré, G. (2022). Artificial intelligence and the conduct of literature reviews. Journal of Information Technology, 37(2), 209–226. https://doi.org/10.1177/02683962211048201

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly, 26*(2), xiii–xxiii. https://doi.org/10.2307/4132319

Yang, Y., Zhang, K., & Fan, Y. (2023). sDTM: A supervised Bayesian deep topic model for text analytics. Information Systems Research, 34(1), 137–156. https://doi.org/10.1287/isre.2022.1124