© *Academy of Management Annals*
2025, Vol. 19, No. 2, 435-439.
https://doi.org/10.5465/annals.2025.0138

FROM THE EDITORS

LARGE LANGUAGE MODELS AS TOOLS FOR INTEGRATIVE REVIEWS[1]

STEFAN THAU
INSEAD

RIITTA KATILA
Stanford University

The scholarly community has witnessed an explosion of interest in large language models (LLMs)—that is, AI systems trained on vast corpora of text that can generate, interpret, and analyze text in response to human prompts. In this editorial, we examine the potential of LLMs to augment management scholars' ability to review literature—to characterize how fields grow, where they shift, and which contributions stand out as influential turning points.

Integrative reviews aim to synthesize diverse literatures. They surface conceptual tensions, identify overarching patterns, and offer conceptual clarity in fields marked by fragmentation (Cronin & George, 2023). Yet, doing this well is hard. Scholars are limited by bounded attention spans, partial recall, and limited interdisciplinary fluency. Our manual coding techniques struggle to keep pace with the rapid expansion of literature across increasingly interconnected fields. Here, LLMs offer something genuinely new. They are not static tools, but collaborative instruments. They can help scale the work of integration and, thanks to the "human API" (i.e., the verbal prompts), LLMs have low barriers to entry.

In this editorial, we ask whether LLMs can serve as *co-scientists*. We do *not* propose that LLMs can replace human expert judgment, but rather suggest that they can assist scholars, who are ultimately responsible for the increasingly complex task of integrative synthesis.

To get a sense of what LLMs can do, consider an analogy to the statistical tools management researchers use to make sense of numerical data. Much like statistical models help scholars filter out signals from noise, LLMs can help sift through large volumes of text to find what matters. Where a correlation matrix reveals relationships among variables, LLMs can surface conceptual patterns. They can identify recurring themes and point to blind spots, as well as identify constructs that go hand in hand, both within and across bodies of literature. In addition, just as regression models help scholars uncover associations and temporal sequences among constructs, LLMs can help show how ideas emerge, evolve, and influence one another, and so help reconstruct intellectual timelines and trace theoretical lineages.

Of course, like all methods, LLMs have limitations. Consider regression models, for example. They can help us understand causal paths, but they can also mislead if poorly designed. Similarly, scholars must use LLMs with great care, with a clear understanding of both their promise and their limitations. After each

LLM-assisted analysis of text, for example, it is critical to explicitly ask the LLM about the assumptions, methods, and code used to produce its text-based inferences. Moreover, because an LLM's analysis depends on the content it sees and is trained on, scholars are responsible for curating the source material. Until open access is widespread, scholars must ensure that a conceptually relevant and tractable set of papers is part of LLM's text corpus (see Appendix A for an example of LLM-assisted analysis of text). In this collaborative workflow, the human remains the integrator; the LLM becomes a co-scientist.

In what follows, we explore how LLMs can support at different stages of the integrative review process —from initial scoping and thematic clustering to temporal mapping and synthesis. With their ability to analyze text to track patterns, cluster themes, identify emerging topics, and score sentiment, we suggest ways in which LLMs could serve as co-scientists. Our goal is not to prescribe a rigid method, but to showcase possibilities; to view LLMs as evolving tools that assist with human scholarly integration.

THE DESCRIPTIVES OF LLMS: THEMATIC LANDSCAPE MAPPING

As a starting point, consider an analogy to the role of variable descriptives and correlation tables in quantitative research. Descriptives report the range and distribution of variables, while correlations indicate the strength and direction of associations between them. In this way, descriptives and correlations offer a structured summary of a dataset. Similarly, when applied to literature corpora, LLMs can help researchers identify the range of topics under investigation, quantify where scholarly attention is concentrated, and surface conceptual linkages between the topics.

First, LLMs can help surface the range of constructs and phenomena studied across a field (or across fields) by generating unsupervised topic clusters. This is much like identifying the spread of variables in a dataset. Second, LLMs can highlight central tendencies and variance in scholarly focus, using frequencies and distributions to show which topics dominate and which are underexplored or outliers. Third, they can model relationships among constructs, and publications, using semantic similarity (based on word usage patterns) or co-citations, producing outputs that resemble conceptual "correlation matrices" between ideas. Together, these analyses can help scholars assess the thematic structure of a field and support clearer conclusions about its shape.

For example, in a review of research on organizational culture, an LLM might identify frequently recurring constructs like psychological safety, norms, identity, and values, and both count and visualize how tightly these constructs cluster across articles (i.e., how often and where they are reproduced). Conversely, this analysis may identify topics that appear rarely or only in niche journals, like organizational culture in crisis management—highlighting a promising gap for future research. Similarly, the LLM-assisted analysis may uncover that psychological safety and gender are each frequently associated with a third, often overlooked, construct that prompts the scholar to suggest a new path for future work.

Naturally, researcher judgment remains essential; but when carefully documented and designed with reproducibility in mind, LLM-assisted analyses can offer a more nuanced and more comprehensive characterization of past research than what most humans can accomplish alone.

The value of LLM-assisted thematic maps also depends on how scholars generate and report them. Correlation tables are interpretable only when readers know how variables were measured and samples selected. Analogously, we need to document the LLM's corpus and the method. For example: What literature and what disciplines were included, and why? How were the topic areas defined and validated? Without transparency, the output risks becoming a black box that reproduces the same imprecision and opaqueness

that Schweinsberg, Thau, and Pillutla (2023) critique in conventional literature reviews. However, when paired with careful documentation and precise descriptions, LLMs can serve to enhance the informational value of integrative reviews by making their empirical inferences more comprehensive, and potentially more nuanced.

THE REGRESSION TABLE OF LLMS: TEMPORAL MAPPINGS

LLMs can also help trace how ideas evolve over time, and so help us understand an intellectual trajectory of a field. Again, an analogy from quantitative research is regression tables. In a regression table, researchers impose structure on numerical data by testing causal paths and estimating prediction strength. Similarly, LLMs can help bring structure to a body of literature by revealing how ideas emerged and evolved over time.

For instance, scholars can use an LLM to analyze the language used in papers together with metadata such as publication dates to map how ideas developed over time. LLMs can help identify which concepts emerged first, how their meaning and empirical operationalizations have shifted, and which new topics they helped shape. Such timelines offer structured views of how ideas diverge or converge across papers. By tracing shifts in language, theoretical framing, or methodological emphasis (see also Katila and Ahuja, 2002), LLMs can help scholars detect turning points in a field—such as the introduction of a new construct, diffusion of an idea across disciplinary boundaries (Katila, 2002), or the rise of an entirely new paradigm.

Another promising feature of LLMs is that when trained properly, they can help scholars identify the causal structure of verbally expressed theories. With refined prompts, LLMs can detect core causal topologies (e.g., direct links, mediators, moderators) and classify papers accordingly. This capability assists scholars to group studies not just by topic but also by causal logic: Which group of (empirical vs. theoretical) papers advocates for moderation, reverse causality, or emphasizes mediation? In this way, LLMs can help identify conflicting assumptions in different subfields and surface theoretical tensions that are key to integrative reviews.

As with any modeling tool, the usefulness depends on clear documentation of the corpus, coding decisions, and interpretive logic. However, when done with care—including by making both the input and the analysis process tractable—this approach can lend temporal structure and broader reach to reviews that otherwise risk remaining impressionistic and underspecified. In other words, LLMs have the potential to assist humans with integrating a larger amount of scholarly material than humans can integrate alone.

HYPOTHESIS TESTING WITH LLMS: TESTING PATTERNS AND CLAIMS

The third useful analogy we propose is between LLM-assisted reviews and hypothesis testing. Hypotheses are formulated about relationships between variables—whether one predicts another, whether the relationship changes over time, or whether it holds across contexts. Similarly, LLMs can be prompted to simulate hypothesis-like inquiries within a body of literature. For example, one might ask: Has theory *X* been more frequently associated with outcome *Y* in one field versus another? How strong are these outcomes, sample sizes, and designs? These are not merely retrieval tasks; they are inferential exercises, and scholars can direct LLMs to carry them out systematically across hundreds of papers.

This way, the role of the review shifts from a retrospective summary to a form of empirical inquiry. LLMs can generate tentative "answers" to meta-level questions about the literature—answers that can be evaluated, refined, or contested by human judgment. Just as exploratory factor analysis helps reveal latent structures in quantitative data, LLMs can help surface hidden patterns in the literature—patterns that

generate new hypotheses rather than simply summarize existing ones.

For example, an LLM might reveal that studies linking psychological safety to team performance disproportionately cite a small cluster of foundational texts, suggesting a potential citation bottleneck or unexamined assumptions in that subfield, or that the relationships have all been supported in some empirical contexts but not in others. Another use case is identifying constructs that appear more frequently in some parts of the manuscript (e.g., theory building) than others (e.g., methods). In the future, rapidly evolving LLMs may even support co-predicting where the field is headed or help generate preliminary ideas for future research that scholars can then refine through expert judgment.

Of course, as with any LLM-assisted analysis, scholars must verify and refine the end results with their own expertise. Still, the use of LLMs in this way opens new possibilities for how management scholars can synthesize literature for integrative reviews, and especially the relationships that have been envisioned across large and fragmented bodies of research.

By learning to use these methods now, we lay the foundation to explore less tractable—but potentially more unexpected and generative—patterns and insights that LLMs may help uncover in the future.

Taken together, we propose that when embedded in a transparent review methodology, LLM-assisted analyses need not be seen as black-box judgments. Instead, they function as structured inferences grounded in text data and subject to scholarly scrutiny. They offer a means of interrogating the literature's internal logic—its conceptual clustering, rhetorical tendencies, and implicit associations—in ways that extend beyond human memory or manual coding. Used responsibly, this approach opens space for a more analytical, interrogative mode of synthesis, one in which reviews not only map what is known but actively question and document how that knowledge has been constructed.

FOOTNOTE

1. In this article, we used ChatGPT 4.0 and 4.5 for editing the text and for the proof of concept. Please see "AOM Artificial Intelligence (AI) Policy" for general guidelines using artificial intelligence: https://aom.org/research/publishing-with-aom/aom-artificial-intelligence-policy

REFERENCES

Cronin, M. A., & George, E. 2023. The why and how of the integrative review. ***Organizational Research Methods,*** 26: 168–192.

Katila, R. 2002. New product search over time: Past ideas in their prime? ***Academy of Management Journal,*** 45: 995–1010.

Katila, R., & Ahuja, G. 2002. Something old, something new: A longitudinal study of search behavior and new product introduction. ***Academy of Management Journal,*** 45: 1183–1194.

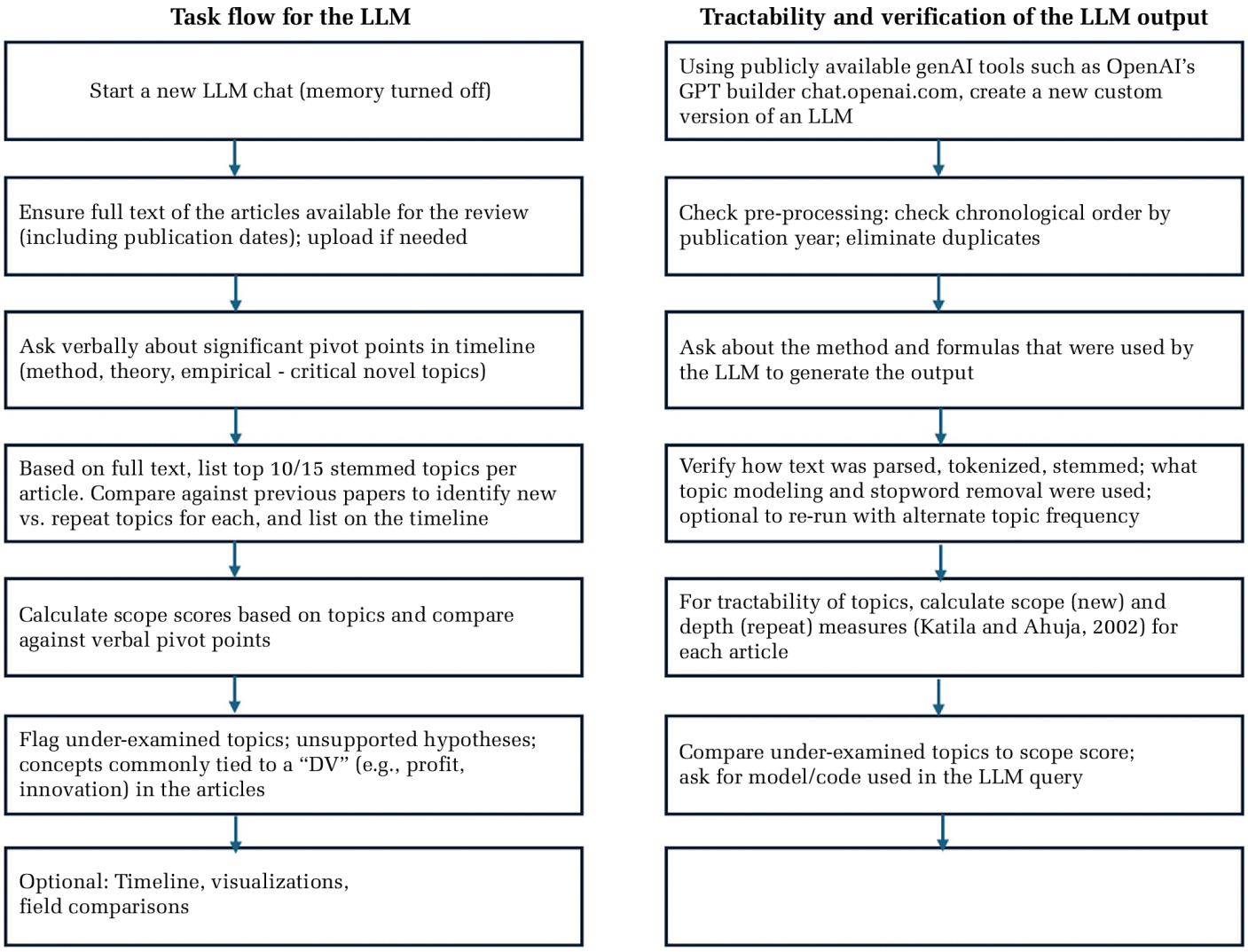
Schweinsberg, M., Thau, S., & Pillutla, M. 2023. Problem validity in primary research: Precision and transparency in characterizing past knowledge. ***Perspectives on Psychological Science,*** 18: 1230–1243.

APPENDIX A

TABLE A1

Proof of concept for using a large language model (LLM) to identify pivot points in a literature for an integrative review. The left flowchart outlines the "human API", i.e., a sequence of prompts for the LLM. The right flowchart outlines how the tractability of the LLM's output can be assessed through follow-up queries at each step. The input to the process can be full text of the papers to be reviewed (or their vectorized versions in the future), or if training data verifiably includes the relevant scholarly articles, it can be

directly used without uploading any material. (Table view)



Sample Prompts

I am a researcher analyzing academic literatures. I upload research papers to this GPT to help me surface important inflection points, shifts, and emerging themes in a field. I value deep, insightful analysis, not just summaries, and I want the GPT to infer and propose plausible pivot points even when they are only partially visible.

I'm uploading the papers now.

Analyze uploaded papers to identify pivot points — major shifts in theories, methods, core questions, or empirical settings. Infer pivots even if not explicit by spotting emerging patterns across papers. For each pivot, give: (1) a short name, (2) a 2–4 sentence description of the shift, (3) key papers involved, and (4) supporting evidence (quotes or synthesized reasoning). Use hedging language ("suggests," "points toward") when necessary. Prefer multiple small pivots over forcing a grand one. Be analytical, insightful, and grounded in the uploaded materials.

Can you cite specific papers, i.e. anchor each pivot point with specific paper(s) from the upload

Extract the top 15 stemmed topics from the full text of the uploaded papers using a combination of topic modelling techniques.

Can you give me exact details of the topic modelling that you used.

I'd like to understand how much each paper is exploring new topics relative to the topics this set of uploaded papers has explored previously. You would first need to organize these papers chronologically.

Can you list for each paper, the topics, new topics vs repeat topics.

I'm going to describe a formula to describe pivot points as "exploration". Are you able to use the topics as raw material to calculate my formula. For each article, give me the top topics, and count how many times those topics have appeared in previous articles (chronologically earlier papers in the upload). Topics that rarely or never appeared before = High Exploration. Topics that frequently appeared before = Low Exploration/High Exploitation.

Can you use my formula to suggest pivots, and then add any other additional pivots that your own analysis suggests.

Can you add under-examined topics from earlier papers that have not been examined deeply in subsequent papers and remain thus open for investigation.