Epistemic framings in science

Charting scientific knowledge with embeddings and LLMs

Christophe Malaterre and Francis Lareau

1. Introduction

Despite the growing attention paid to the diversity of scientific practices, relatively little is known about how scientists conceptualize and articulate the very nature of knowledge in different research fields. Philosophers of science have long debated key epistemic constructs—such as theory, model, hypothesis, explanation, and understanding—and have emphasized the roles of justification and epistemic virtues like simplicity, robustness, and predictive power. Yet, these discussions often rest on abstract conceptual analysis or a narrow set of case studies, rarely examining how such epistemic notions are used in scientific discourse at scale (but see McPhetres et al., 2021; Mizrahi and Dickinson, 2022; Noichl, 2023).

Our recent corpus-driven research begins to fill this gap by examining how scientists actually deploy a vast array of epistemic concepts in their published work (Malaterre and Léonard, 2023, 2024). Using a large biomedical corpus, we analyzed the distribution and association of dozens of epistemic markers (e.g., "theory", "model", "explanation") revealing a clear diversity of usage patterns across subfields. These differences suggest not only semantic variation but deeper conceptual distinctions—what we refer to as "epistemic framings": the ways in which scientists frame what counts as knowledge, how it is justified, and which epistemic goals it serves.

Here we explore how recent advances in natural language processing—particularly embedding models and large language models (LLMs)—can contribute to the systematic investigation of epistemic framings. We argue that these tools offer powerful methods for detecting, classifying, and mapping epistemic concepts across disciplines and historical periods, thus enabling a more empirically grounded account of how scientific knowledge is conceptualized and communicated.

2. A corpus study: epistemic concepts across biomedical subfields

Our initial empirical work focused on a biomedical corpus of 73,771 full-text research articles, pre-processed with lemmatization and sorted into seven disciplinary clusters via topic modeling and k-means clustering. Our aim was to assess how epistemic concepts are deployed across subfields—how "model," "explanation," or "mechanism," for instance, figure differently in neuroscience, epidemiology, or molecular biology (Malaterre and Léonard, 2024).

To start, we developed a curated list of 211 epistemic markers through two complementary strategies. First, we expanded from six seed terms—"explanation," "understanding," "prediction," "model," "theory," and "mechanism"—by extracting their most frequent co-occurring terms across the entire corpus and within each cluster. This yielded 202 high-frequency terms, 52 of which were epistemic in nature. Second, we analyzed the most frequent terms in the overall corpus (appearing in at least 1,000 documents), supplemented by thesaurus-based synonym expansion. The overall 211 terms were manually grouped into 61 epistemic "semantic fields" (e.g., hypothesis, confirmation, plausibility) and 18 disciplinary ones (e.g., disease, molecular, behavior).

We then quantified the presence of each semantic field in each document and computed averages by disciplinary cluster. Additionally, we calculated paragraph-level co-occurrence correlations between all epistemic fields, producing a 79×79 correlation matrix per disciplinary cluster. Louvain community detection on these correlation networks allowed us to map discipline-specific patterns of epistemic co-occurrence.

Key findings showed that different disciplinary domains privilege different epistemic objects—e.g., models in neuroscience, mechanisms in molecular biology—and epistemic virtues—e.g., plausibility in clinical research vs. accuracy in genomics. These terminological "epistemic maps" reveal that scientific disciplines structure and prioritize epistemic concepts in distinct ways. In turn, this supports the idea that scientific knowledge is not monolithic, but plural and context sensitive. We call these structured configurations of epistemic emphasis "epistemic framings."

3. Using LLMs to classify epistemic content

Large language models (LLMs) offer promising tools for extending this investigation beyond fixed lexicons. One avenue we are experimenting with is the contextual classification of epistemic statements. LLMs can be prompted with individual sentences or paragraphs and asked whether the epistemic content expresses certainty, uncertainty, evaluative judgement, or justification. These classifications could be used to analyze the epistemic orientations of scientists as expressed in their written productions.

However, the sheer scale of large scientific corpora limits this approach in practice. For instance, analyzing 100 million sentences with a model like GPT-4-turbo would require weeks of compute time and a substantial cost. This makes full-coverage classification with current LLM APIs impractical, though this may change in the future as LLM cost and compute time drop down.

Instead, LLMs are probably best deployed in a hybrid architecture, at least for now. Small samples of sentences can first be annotated via LLMs to create a training sets with which lighter classifiers can then be trained to label the rest of the corpus. The resulting classification can then be used as a heuristic to identify a variety of epistemic framings, and investigate nuances depending on disciplinary domains or time.

Ultimately, while LLMs cannot solve the entire epistemic classification task at scale, they can significantly enhance the sensitivity and coverage of framing analyses by providing contextual judgments, expanding lexicons, and improving interpretability when prompted accordingly on relatively small sets of documents.

4. Refining epistemic markers with embeddings and syntactic patterns

One limitation of our earlier study lies in its reliance on a manually curated list of epistemic terms. While conceptually grounded, the list may miss field-specific phrases, idiomatic expressions, or emergent terms. Embedding models like SciBERT, or prompting LLMs directly, allow for semantic expansion: generating synonyms and near-synonyms of known epistemic markers (e.g., "tentative explanation," "broadly consistent," "likely mechanism").

We also propose to enhance the lexicon by identifying multiword expressions (MWEs) with epistemic relevance. These can be discovered by parsing sentences that contain known epistemic terms and extracting syntactic patterns such as modal constructions (e.g., "may explain," "is thought to be," "cannot confirm"). Dependency parsing tools like spaCy or Stanza can isolate these structures.

By clustering embeddings of parsed dependency structures (e.g., using HDBSCAN), we can identify families of expressions that share structural and semantic characteristics. Clusters enriched in epistemic content can yield additional MWEs not present in the original list.

This syntactic-semantic approach captures not just individual terms but the broader constructions in which epistemic meanings are embedded, offering a more nuanced view of how scientific claims are hedged, justified, or asserted.

5. Topic modeling and disciplinary structure

Understanding epistemic framings requires contextualizing them within the thematic and disciplinary structures of science. LLM-based topic modeling offers a way to do this. One approach is to embed documents (e.g., abstracts or full texts) using sentence-transformers like SBERT or domain-specific variants (e.g., BioBERT). Dimensionality reduction techniques like UMAP, followed by clustering algorithms like HDBSCAN, can then identify coherent topic clusters.

Once these clusters are formed, characteristic terms can be extracted using class-based TF-IDF (c-TF-IDF), F-measure, or LLM-based summarization. LLMs can also be used to evaluate different models (Stammbach et al., 2023), and can assist in generating descriptive labels for these clusters, increasing interpretability.

Tools like BERTopic automate this pipeline (Grootendorst, 2022), combining embeddings, clustering, and topic extraction. However, trade-offs exist. Embedding-based topic modeling can struggle with long documents due to window size limitations. It may also underperform compared to classical statistical methods (e.g., LDA (Blei et al., 2003)) in terms of topic coherence or granularity (Lamirel et al., 2024). Thus, hybrid approaches—combining embeddings for clustering and statistical term scoring for extracting key topic terms—can be effective.

Note that the granularity of the textual unit matters (see Lareau and Malaterre, 2026): full texts may blur topic boundaries, while paragraphs or abstracts offer more focused representations. Choosing the right scale depends on the epistemic resolution sought.

Ultimately, embedding-based topic modeling allows us to situate epistemic framings within disciplinary contexts, supporting more precise analyses of how knowledge is construed differently across research areas.

6. Toward a theory of epistemic framings

Traditionally, philosophical analysis has treated epistemic concepts—such as "theory," "explanation," or "evidence"—in isolation, seeking precise definitions and logical relationships. But science is practiced in varied contexts (Knorr-Cetina, 1999; Chang, 2012; Ankeny and Leonelli, 2016), and these concepts vary in usage and may acquire different meanings depending on epistemic aims, methodological traditions, and disciplinary norms.

The idea of epistemic framing offers a way to theorize this variability. An epistemic framing is not just a preference for one term over another, but a patterned configuration of epistemic priorities, conceptual tools, justificatory practices, and linguistic forms that collectively express how a field construes knowledge. Prior work we conducted (Malaterre and Léonard, 2024) shows the relevance of this notion to describe how epistemic concepts function differently across biomedical subfields.

Why does this matter? First, it supports a version of scientific pluralism: the idea that science is not governed by a single set of epistemic norms (Longino, 2002). Second, it aligns with epistemic contextualism, suggesting that key scientific concepts have context-sensitive meanings. Third, it opens the door to data-driven philosophy of science: by integrating empirical tools with conceptual analysis, we can better understand the landscape of scientific knowledge production (Ramsey and De Block, 2022).

By analyzing the discourse of science at scale, we can move beyond abstract theorizing and map the real-world structures through which scientists articulate and justify knowledge. Investigating epistemic framings with quantitative text-mining methods complements traditional conceptual work in the philosophy of science. It brings new tools to long-standing debates and provides a platform for interdisciplinary collaboration between philosophers, computational linguists, and practicing scientists.

While current LLM-based methods still face scalability and interpretability challenges, the promise is clear: with refined tools and curated corpora, we can chart the evolving epistemic landscapes of science—revealing how knowledge is conceptualized, justified, and valued in different corners of scientific practice.[1]

Acknowledgements

F.L. acknowledges funding from Canada Social Sciences and Humanities Research Council (Postdoctoral Fellowships 756–2024-0557, Grant 430–2018-00899). C.M. acknowledges funding from Canada Social Sciences and Humanities Research Council (Grant 430–2018-00899) and Canada Research Chairs (CRC-950-230795). The authors thank Arno Simons, Adrian Wüthrich, Michael Zichert and Gerd Graßhoff for organizing the *LLMs in HPS* workshop (TU Berlin, 2025) and inviting them to contribute to this volume.

References

Ankeny RA and Leonelli S (2016) Repertoires: A post-Kuhnian perspective on scientific change and collaborative research. *Studies in History and Philosophy of Science Part A* 60: 18–28.

Blei M, Ng AY and Jordan MI (2003) Latent dirichlet allocation. *Journal of Machine Learning Research* 3(1): 993–1022.

Chang H (2012) *Is Water H2O?: Evidence, Realism and Pluralism*. Dordrecht: Springer Netherlands.

Grootendorst M (2022) BERTopic: Neural topic modeling with a class-based TF-IDF procedure (arXiv:2203.05794). *arXiv*. DOI: 10.48550/arXiv.2203.05794

Knorr-Cetina K (1999) *Epistemic Cultures: How the Sciences Make Knowledge*. Cambridge (MA): Harvard University Press.

Lamirel J-C, Lareau F and Malaterre C (2024) CFMf topic-model: Comparison with LDA and Top2Vec. *Scientometrics* 129(10): 6387–6405.

Lareau F and Malaterre C (2026) When less is more? Topic model performance across title, abstract, and full-text corpora. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-4.

Longino HE (2002) *The fate of knowledge*. Princeton (NJ): Princeton University Press.

Malaterre C and Léonard M (2023) Charting the Territories of Epistemic Concepts in the Practice of Science: A Text-Mining Approach. *The British Journal for the Philosophy of Science*. Epub ahead of print. DOI: 10.1086/725656

Malaterre C and Léonard M (2024) Epistemic markers in the scientific discourse. *Philosophy of Science* 91(1): 151–174.

1 This chapter was written with support from large language models (LLMs). All model-generated text was reviewed and, where necessary, rewritten by the authors, who remain fully responsible for the final version. For details on the use of LLMs in this volume, see the statement in the volume's introduction.

McPhetres J, Albayrak-Aydemir N, Mende AB, et al. (2021) A decade of theory as reflected in Psychological Science (2009–2019). *PLOS ONE* 16(3): e0247986.

Mizrahi M and Dickinson MA (2022) Philosophical reasoning about science: A quantitative, digital study. *Synthese* 200(2): 138.

Noichl M (2023) How localized are computational templates? A machine learning approach. *Synthese* 201(3): 107.

Ramsey G and De Block A (2022) *The Dynamics of Science: Computational Frontiers in History and Philosophy of Science.* Pittsburgh (PA): University of Pittsburgh Press.

Stammbach D, Zouhar V, Hoyle A, et al (2023) Revisiting Automated Topic Model Evaluation with Large Language Models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* 9348–9357.