

CorpusLens: LLMs as Data Analysis Facilitators in the Social, Humanistic, and Language Sciences

Blake Vente^{*1}, Nicholas Deas^{*1}, Desmond Patton²,
Shana Kleiner², Jessica A. Grieser³, Kathleen McKeown¹,

¹Columbia University, Department of Computer Science

²University of Pennsylvania, School of Social Policy and Practice,
Annenberg School for Communications

³University of Michigan, Department of Linguistics

rv2459@columbia.edu {ndeas, kathy}@cs.columbia.edu

Abstract

In our interdisciplinary, collaborative research, we have found that humanities and social science researchers would like to analyze data underlying LLMs as well as study how language tools (e.g., emotion detection) work on their data. Existing tools, while potentially enabling, are often unintuitive for such researchers or do not provide them enough agency over how the analysis is conducted. In this paper, we introduce **CorpusLens**, an LLM-based tool to aid social, humanistic, and language science researchers in exploratory textual data analysis. We present the design of CorpusLens as well as a user study showing that researchers found CorpusLens to be more helpful than prior tools they had used, while also rating it as intuitive, easy to use, transparent and enabling of novel analyses.¹

1 Introduction

In a long-standing collaboration among social work researchers, sociolinguists, and computer scientists, our team has studied large language models' (LLMs) understanding of African American Language (AAL) (Deas et al., 2023; Kleiner et al., 2024; Grieser et al., 2023; Deas et al., 2024, 2025; McKeown et al., 2026) alongside digital expressions of Black grief (Patton et al., 2025a, 2024, 2025b). Throughout our collaboration, language technologies have been both the subject of critical study (e.g., identifying the ways in which LLMs fail to comprehend AAL) as well as the tools we leveraged to analyze large-scale textual data to investigate social, humanistic, and linguistic questions (e.g., surfacing themes in discussions of grief online).

^{*}denotes equal contribution

¹We make our code available at <https://github.com/Columbia-NLP-Research-Lab/corpuslens/>, and the tool can be accessed at <https://corpuslens.com/>.

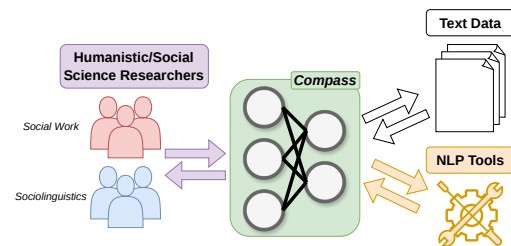


Figure 1: Overview of the CorpusLens system. An LLM data analysis agent (COMPASS) facilitates social, humanistic, and language science researchers' data analyses, serving as an interface to the underlying textual data as well as NLP tools for text analysis.

Our follow-up studies to our work showing that LLMs struggle to understand AAL (Deas et al., 2023) have, in part, focused on analysis of data to understand *why* they struggle. In particular, we have worked toward identifying particular constructions that were most difficult for LLMs (Grieser et al., 2023), diving deeper into LLMs' understanding of highly performative use of AAL online (Kleiner et al., 2024), and quantitatively analyzing the representation of AAL in large-scale pretraining corpora (Deas et al., 2025). In all cases, we have seen the need for insights from the different disciplines that make up our team: artificial intelligence (AI) focusing on the use and development of analysis tools, sociolinguistics focused on AAL and its variation, and social work research contextualizing AAL use and the impacts of LLMs' failures.

Bridging distinct disciplines to pursue these questions, however, also presents unique challenges. The general aim of the aforementioned studies was to ask linguistic and humanistic questions about large language corpora. While tools such as *What's in My Big Data?* (WIMBD; Elazar et al. 2024) and Infinigram (Liu et al., 2024; Xu et al., 2025) enable efficient search of large corpora,

they do not lend themselves to more complex analyses (e.g., grammatical patterns, semantic features) that may yield even more novel insights into the data. Furthermore, the tools to conduct these or related analyses often exist behind high technical barriers to their use. Therefore, there is a clear need to bridge the gap between these AI tools and the linguistic and humanistic researchers that may benefit from them.

Our overall aim is therefore to leverage LLMs as an *interface* between researchers, textual data, and language tools for analyzing such data (illustrated in Figure 1). In beginning to develop an interface for exploratory data analysis, we quickly found that mimicking traditional graphical interfaces was un-intuitive for social, humanistic, and social science experts. Through feedback from informal focus studies, we converged instead on LLMs as *facilitators* of data analysis. By facilitators, we mean that the model is positioned such that it can provide more intuitive access to analysis tools, but also such that decision-making and agency remain with the researcher.

Our main contributions are as follows. Building upon an initial design found to be unintuitive (Section 3.1), we present an improved, **co-designed LLM-based interface—CorpusLens**—to better ensure the system is usable and enabling to researchers (Section 3.2). Domain experts positively evaluate the tool and rate it as **more helpful on average than alternative software** for complex text analyses. Finally, based on this evaluation, we discuss **opportunities and challenges** for LLMs in this setting (Section 5).

2 Related Work

LLMs and Research Processes. LLMs have increasingly been used to aid researchers in various stages of research. For example, LLMs have been investigated for generating research hypotheses (e.g., Zhou et al. 2024; Liu et al. 2025a) and conducting literature reviews (Tang et al., 2025; Wang et al., 2026) such as AI2’s Asta². In particular, LLMs and agents have recently been leveraged for automating data analysis, documentation, and interpretation (Hong et al., 2025). We focus on this aspect of the research pipeline, but rather than studying LLMs for annotating data or conducting independent analysis, we study LLMs as *facilitators*, interfacing between researchers, the data to

be analyzed, and the tools necessary to do so.

LLMs and Humanistic/Social Science. Significant progress has been made toward leveraging AI for scientific discovery in, for example, the physical sciences (e.g., modeling protein and DNA sequences), mathematics (e.g., automated theorem proving), and engineering (e.g., *inverse design* problems) (van Dijk and Vrkc, 2026; Wang et al., 2023). LLMs have also, however, been studied in the context of scaling deductive coding and data labeling (e.g., Chew et al. 2023) as well as inductive coding processes (Subbiah et al., 2026) that are common in humanistic and social science fields. In linguistics, recent work has also assessed LLMs’ capabilities in meta-linguistic reasoning (Yang et al., 2025). Many of these studies, however, support that LLM-based approaches can lack transparency and introduce unique biases into the analysis and interpretation process.

3 System Design

Recall that our primary aim is to develop a system enabling researchers in the social, humanistic, and language sciences to leverage NLP-based data analysis methods (e.g., emotion classifiers, dialect classifiers, etc.). Accordingly, we focus on tools related to thematic and linguistic analysis of text. We additionally prioritize the following core values in designing the system:

1. **transparency**—users should actively be made clearly aware of how analyses are being conducted with provenance details available;
2. **usability**—the interface should be intuitive to use without requiring familiarity with or deep technical knowledge of NLP methods;
3. **user agency**—analysis decisions should be driven by users rather than the LLM; and
4. **interactivity**—data exploration should be interactive and iteratively refined rather than a set of static decisions.

We begin by discussing an initial structured version of the interface. We then discuss the transition to a final LLM-based system based on discussions among collaborators.

3.1 Structured Interface

To design CorpusLens, we explored several versions, iteratively refining our design based on feedback from the sociolinguists and social work researchers on our team. In the initial version, users

²<https://asta.allen.ai/>

were presented with stages of analysis (interface shown in [Appendix A](#)). In the first stage, users prepared the data in the system, selecting which corpus to analyze, the number of records to sample, and various preprocessing steps (e.g., generating text embeddings or predictions of analysis models). Following this, users were provided options to augment and filter the data sample based on numerical (e.g., sentiment classification scores) and/or qualitative (e.g., keyword search) measures. Finally, selected plots were generated depending on visualization choices (e.g., UMAP plots).

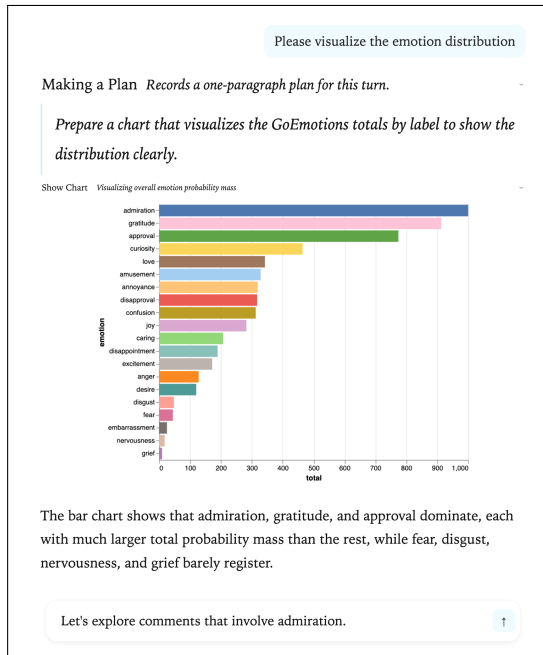


Figure 2: Sample compressed screenshot of *LLM-based interface* of CorpusLens. The interface mimics a typical LLM dialogue window, including inline data visualizations, tables, and structured requests for user input.

This interface was created to expose users to data analysis choices, such as filtering, in a structured way while minimizing the need for highly technical decisions regarding specific analysis methods. Through discussions with our collaborators in social work and sociolinguistics, we found, however, that this initial interface had a few key limitations. First, the initial tool was unintuitive, placing a high burden on users to make many decisions before receiving results. Additionally, the scaffolded stages make interactive exploration difficult, with small changes requiring rerunning earlier stages. Finally, the static interface is unable to adapt to users' needs (e.g., those that may wish to see underlying SQL queries vs. those that would not find the underlying queries useful).

3.2 LLM-Based Interface

Due to these limitations, our discussions culminated in an LLM-based interface ([Figure 2](#)) to lower barriers to using the available tools, enable interactive analyses, and dynamically adapt to different users.

LLM Facilitator Agent. The core component of the interface is an LLM agent, which we call COMPASS. COMPASS has three primary functions: 1) interpret user queries, 2) carry out analyses, and 3) display results. In other words, COMPASS acts as a natural language interface to a set of analysis tools similarly to Text-to-SQL approaches interfacing with databases (e.g., [Eyal et al. 2023](#)). COMPASS, however, enables iteratively refining queries, can adapt to user feedback, and is provided tools for visualizing data and modifying the interface to facilitate data exploration.

Database. All data being analyzed, tools available to COMPASS, and additional information (e.g., emotion/sentiment labels) are centrally stored in a DuckDB database ([Mühleisen and Raasveldt, 2026](#)). For the database, COMPASS is restricted only to SQL queries; generally, COMPASS is allowed to perform operations on the database tables through user-defined functions (UDFs) available in DuckDB (e.g., a UDF for adding sentiment predictions). These constraints are implemented toward ensuring that COMPASS remains within the scope of the users' request and carrying out analyses rather than performing additional analyses or freely modifying the data.

Agent Tools. The COMPASS agent is provided a set of tools for processing data, presenting results, and modifying the interface. Through *data analysis* tools, the agent is able to supplement data with additional information (e.g., sentiment and emotion labels, dense embeddings) and filter data samples in order to address user queries. All tools are defined and described in a database table accessible to the COMPASS agent, allowing for easily incorporating additional tools for use by the agent as well as displaying such details to the user when requested.

Through interface tools, the agent can make adjustments to the interface such as collapsing or un-collapsing prior results or showing intermediary steps of a prior analysis based on user requests. Through *visualization* tools, the agent is capable of displaying interactive plots and charts to the user. Notably, users can select portions of a chart in order to ask targeted questions about a subset of data in a

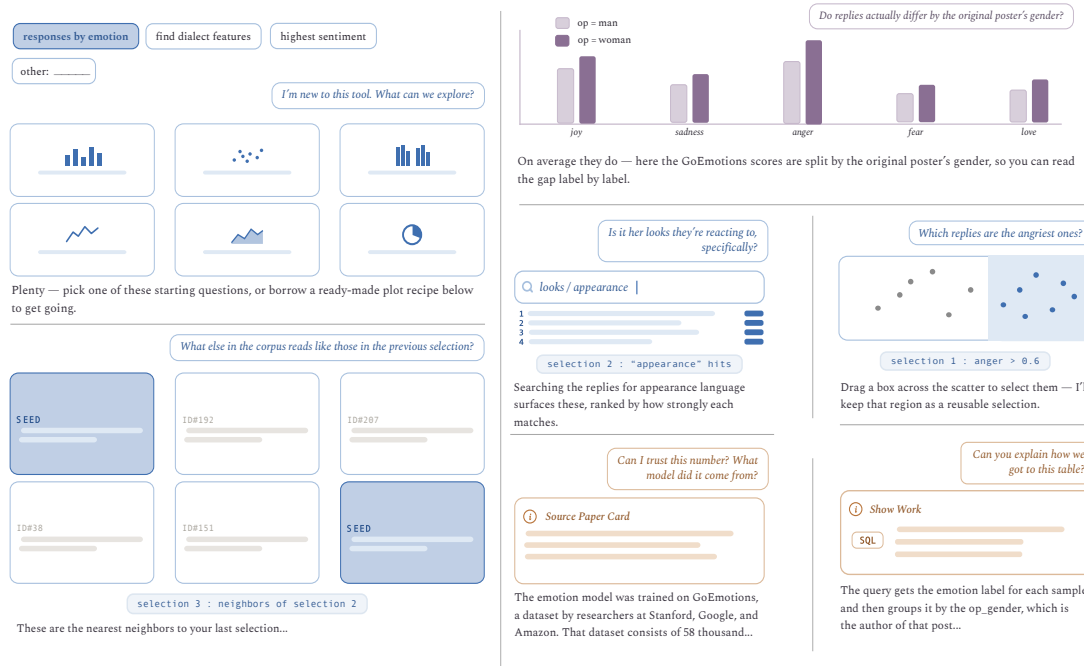


Figure 3: Interactions within a typical CorpusLens session, the initial questions to the user. The figures depict interaction types enabled by the choice of framework. Notably, the assistant can create components with the capacity for direct manipulation by the users.

visualization (e.g., "what is the dominant emotion expressed in *these* posts?" where "*these*" refers to a region in a scatter plot highlighted). Select interactions with the model and available tools are illustrated in Figure 3, and full details on the initial set of tools implemented for this study are included in Appendix C. Notably, CorpusLens is intended to allow research teams to flexibly add tools to the system and adapt it to their own collaborative work.

What happens when there's a query? At the beginning of a CorpusLens session, a pre-defined *introductory prompt* is shown to the researcher. Specifically, the interface asks broad, optional questions about the researchers' intent (e.g., learn about data analysis, explore a dataset), the kinds of questions they are interested in, and preferences on explanations. If responses are provided, they are passed to COMPASS to personalize the session. The session is then user-driven primarily through natural language analysis and visualization queries (e.g., first user message in Figure 4) or other interactions with the interface (e.g., selecting a plot region). Correspondingly, COMPASS can respond to user queries or actions with analysis tool calls, visualization tool calls, or similar interface interactions (e.g., collapsing prior result displays). Additional architectural details are discussed in Appendix B.

4 Domain Expert Evaluation

To evaluate the effectiveness of LLMs as data analysis facilitators in CorpusLens, we conduct a user study with domain experts—social work researchers, communication scholars, and sociolinguists.

4.1 Methods

Implementation Details. In the domain expert evaluation, the COMPASS agent is built on GPT-5.1 Codex ³ over the OpenAI API. For the purposes of the evaluation, the results of analysis tools are precomputed because our focus is on the evaluation of an LLM as an interface.

Domain Experts. To evaluate the system, we work with two research labs that are representative of the intended audience of CorpusLens. First, we work with social work and communication scholars at the University of Pennsylvania; the groups' work broadly centers on computational and social work approaches to understanding violence, joy, and grief. We additionally work with sociolinguists at the University of Michigan focusing on AAL, on-line language, and other intersections of language and identity. For the study, ten researchers use the system to analyze data and complete an evaluation survey.

³<https://chatgpt.com/codex/>

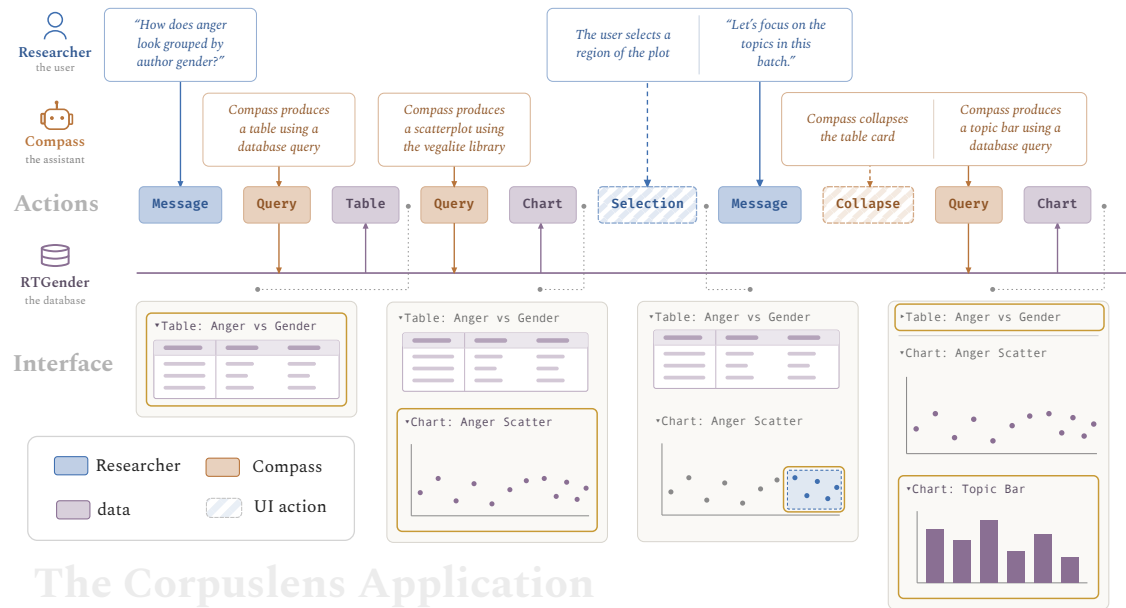


Figure 4: An example multi-turn interaction between a user and CorpusLens analyzing the RTGender dataset (Voigt et al., 2018). Actions performed on or by the system during the **session** are listed in order. As a **researcher** sends messages to the system, **COMPASS** uses tools to conduct analyses on or display visualizations of the **database**. Both the **researcher** and **COMPASS** have a shared ability to **modify** the interface in addition to sending messages.

Procedure. Participants are provided access to the CorpusLens tool, a guide on how to use the tool, and three possible datasets from prior work for potential analyses: RTGender (Voigt et al., 2018), the ChangeMyView corpus (Tan et al., 2016), and the Corpus of Regional African American Language (CORAAL; Farrington and Kendall 2021). These were selected to represent socially and linguistically relevant textual datasets drawn from social media (RTGender and ChangeMyView) and spoken (CORAAL) sources. Each participant is asked to use CorpusLens to freely choose one of these datasets and analyze it for at least 20 minutes over the course of two weeks. Importantly, while this initial evaluation is conducted in a controlled setting, we do not guide participants on what analyses to conduct to better reflect realistic uses of CorpusLens. At the conclusion of the period, participants are asked to complete a survey reflecting on their experience with the tool.

System Evaluation. In the evaluation survey, we include questions focused on our originally stated values. Participants are first asked about their **overall** experience with the tool as a summary measure. We then specifically ask about the **helpfulness** of the tool with respect to common data analysis tools specific to relevant disciplines

(e.g., AntConc⁴, NVivo⁵). Finally, we ask four questions about the extent to which the tool was **intuitive**, **easy-to-use**, **transparent**, and **enabling** of novel analyses. All responses are on 5-point Likert scales (0-4, inclusive), with higher scores representing better evaluation of the system. We additionally ask optional open-ended questions concerning any insights gained from using the tool and comments about the tool and its features. Full study details are included in Appendix D.

4.2 Results

Dimension	Mean Rating
Overall	3.0 ± 0.41
Helpful	2.9 ± 0.54
Intuitive	2.7 ± 0.59
Easy	2.7 ± 0.72
Explain	2.7 ± 0.51
Novelty	2.9 ± 0.35

Table 1: Average scores (0 to 4; scores > 2 are positive ratings) and 95% confidence intervals of the domain expert evaluation.

The results of the domain expert evaluation are

⁴<https://www.laurenceanthony.net/software/antconc/>

⁵<https://lumivero.com/products/nvivo/>

shown in Table 1. Notably, all scores evaluating the system are positive (i.e., > 2), including the overall evaluation of system (~ 3 on average). Most users indicate that CorpusLens is more helpful for the kinds of analysis they conducted than the alternative they use most (e.g., annotation and analysis software like Dedoose). Among the other evaluation dimensions, the intuitiveness and ease of the system are scored lowest, indicating that a natural language interface alone may not fully alleviate barriers to using available tools.

5 Discussion

From participants' open-ended survey responses, we additionally identify opportunities and challenges of LLMs in this domain.

Opportunity: Methodology Provenance and Transparency One of the most salient observations is that methods require comprehensive documentation. When any feature extraction is applied, the method used needs to be visible. Participants wanted to be able to interrogate findings to understand the validity of the inferences, and information about how numerical figures are derived, including source texts. Domain experts appreciated *"the transparency the tool provides users with the plan"* exposed while analyses are calculated. Additional approaches in recent work on provenance such as tracing model outputs to training data (Liu et al., 2025b) or work on science communication (e.g., (Alshomary et al., 2026)) and literature search are promising directions to further provide transparency and clarity on underlying methods.

Challenge: Communicating Capabilities. LLMs and chat-based interfaces are open-ended, enabling a wide array of possible interactions and analyses. Nonetheless, there are practical limits to what users should expect the tool to be able to accomplish within a particular domain. We mitigate this by giving COMPASS access to tools that can demonstrate its capabilities (see Appendix C). Indeed, one participant found that *"the tool was, when prompted, very clear in saying what it can and can't do"* which enabled them to *"pivot [their] queries better."* They also, however, still reported needing *"several attempts to fully understand the tools' purpose and capabilities"* or additional examples *"searches/examples of prompts"* to calibrate expectations. Broadly, effectively communicating capabilities and limitations as well as aligning users' expectations of LLM-based tools with those of the

developers remain important challenges to address.

Opportunity: Incorporating & Adapting to User Feedback. Throughout development and in comments on the final system, domain experts reported a need for the tool to better adapt to their feedback. In one sense, users wanted the ability to correct outputs of the tool and COMPASS agent which would impact future behavior. This capability was incorporated by allowing feedback to be given through interaction with figures. This gap in personalization, however, is also clearly seen in differing responses from domain experts: while one user found that *"the tool often started with very technical explanations that included a lot of statistics and specialized jargon"*, another user *"like[d] that the LLM [could] explain things in laymans terms."* Although COMPASS did not fully adapt to each user, this gap presents an opportunity for recent work on inferring user preferences and personalizing responses in naturalistic styles and settings (e.g., Ashok Kumar et al. 2025; Gao et al. 2024).

Challenge: Data Ethics and Transparency. During the development process, at multiple points in time, concerns were raised about the handling of data. Outside of the controlled setting, researchers would upload their own data to a tool like CorpusLens, which may include sensitive, personal, or otherwise confidential information. Similarly, our collaborators and domain experts involved in the user study expressed concerns about where data and prompts may be sent or stored. In the domain expert evaluation, users also expressed interest in further metadata and documentation of the underlying data sources: *"it would be helpful...to have more information...describing the data and variables..."* These data-centric concerns signal a need for further development of open-source models that can be self-hosted ⁶ (e.g., Groeneveld et al. 2024) as well as data documentation practices (Bender and Friedman, 2018; Gebru et al., 2021). To this end, we provide a version of CorpusLens running entirely within the browser itself, which can be configured to use any Anthropic, OpenRouter, or OpenAI-compatible endpoint, making it usable with local open models, entirely privately.

Opportunity: Enabling Novel Analyses. While the study with domain experts is relatively brief, participants noted novel patterns they were able to find in the provided data using CorpusLens.

⁶That is, run on the researcher's own computer instead of a cloud service

In analyzing CORAAL, one participant noted that they were able to "[learn] about the different regional patterns that emerged from the dataset" through the available tools. Another participant that described their investigation in depth discussed being able to find relationships between emotion, gender, and mentions of sex work in the RTGender corpus (e.g., finding that mentions of sex work are slightly more common in replies addressed to women). They note, however, that the tool surfaced "*many posts on sex, but not necessarily on sex work*" so the researcher still needs to scrutinize discovered patterns and resulting claims, and suggest that the tool is helpful for "*getting an idea of a research question*," with more specific use-cases needing additional tuning of the underlying model.

Opportunity and Challenge: Domain Expertise. A common thread through versions of applications is that models can often lack particular domain knowledge necessary for routine analytical tasks, and need additional intervention or specification to perform well. For example, while the model underlying COMPASS appeared to have a general understanding of the use of *existential it* (e.g., *it is raining*), COMPASS did not fully understand its more specific use in the context of AAL, which would require additional explanation and correction by the user. Overall, while a general purpose language model may provide an easier interface to tools and data, domain-specific models may still be required for more in-depth analyses. This challenge, however, also highlights the opportunity for interdisciplinary work bridging both domain expertise and technical knowledge to pursue new questions aided by computational linguistics tools.

6 Conclusion

This work is a result of a long-standing collaboration among computational linguists, social work researchers, and sociolinguists, intending to address the challenges of bridging disciplines in interdisciplinary work. We introduce CorpusLens, an LLM-as-an-interface tool intended to enable humanistic and social science researchers to leverage NLP tools for exploratory data analysis. Through repeated feedback, we iterated on the tool moving from an originally unintuitive, structured interface to the LLM-based interface discussed in this work. Domain experts positively evaluate CorpusLens, on average reporting it is more helpful than alternative software. Based on the development process and

domain experts' feedback, we additionally discuss opportunities and challenges in LLM-based interfaces for humanistic and social science researchers.

Limitations

We note a few limitations accompanying this work. First, our domain expert evaluation include a fixed set of tools, limiting potential analyses. The tools, however, were selected based on discussions with domain experts and we note that CorpusLens is intended to be easily extensible. Additionally, the domain expert evaluation is limited to datasets relevant to—but not grounded in—users' actual research. To encourage ecological validity, we allow each participant to explore the data however they like and we similarly select datasets through discussion with our collaborators to ensure relevance. Studying tools like CorpusLens in longer term, realistic research settings would be valuable for future work.

Furthermore, we focus on the usability and utility of CorpusLens in this work, but we do not explicitly assess the accuracy of the systems queries and outputs. We rely on a frontier model to avoid significant response quality concerns, but the model may still make errors in matching user intent, correctly accessing the data, and correctly visualizing results. While there are no clear errors or trends from qualitative examination, there are cases of COMPASS producing errored queries due to being asked for charts outside of its capabilities (e.g., producing a geographic map).

7 Acknowledgements

This work was supported in part by National Science Foundation Graduate Research Fellowship DGE-2036197, the Columbia University Provost Diversity Fellowship, and the Columbia School of Engineering and Applied Sciences Presidential Fellowship. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We thank the anonymous reviewers for discussions and feedback on earlier iterations of this work. We also thank Anthony Shaw and other researchers in the UPenn SAFELab for their frequent and helpful feedback on CorpusLens during development.

Ethics

Only artifacts introduced in prior studies and publicly available corpora are used in this work. The

domain expert evaluation in this work was reviewed and approved by Columbia University’s Institutional Review Board (Protocol Number # ACYY0956).

References

- Milad Alshomary, Grace Li, Anubhav Jangra, Yufang Hou, Kathleen McKeown, and Smaranda Muresan. 2026. [Llms as science journalists: Supporting early-stage researchers in communicating their science to the public](#). *Preprint*, arXiv:2601.05821.
- Dimosthenis Antypas, Asahi Ushio, Jose Camacho-Collados, Vitor Silva, Leonardo Neves, and Francesco Barbieri. 2022. [Twitter topic classification](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3386–3400, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Nischal Ashok Kumar, Chau Minh Pham, Mohit Iyyer, and Andrew Lan. 2025. [Whose story is it? personalizing story generation by inferring author styles](#). In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1485–1540, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.
- Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. [TweetEval: Unified benchmark and comparative evaluation for tweet classification](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.
- Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. [SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter](#). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Emily M. Bender and Batya Friedman. 2018. [Data statements for natural language processing: Toward mitigating system bias and enabling better science](#). *Transactions of the Association for Computational Linguistics*, 6:587–604.
- Su Lin Blodgett, Lisa Green, and Brendan O’Connor. 2016. [Demographic dialectal variation in social media: A case study of African-American English](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.
- Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. [Nuanced metrics for measuring unintended bias with real data for text classification](#). *Preprint*, arXiv:1903.04561.
- Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Robert Chew, John Bollenbacher, Michael Wenger, Jessica Speer, and Annice Kim. 2023. [Llm-assisted content analysis: Using large language models to support deductive coding](#). *Preprint*, arXiv:2306.14924.
- Evan Czaplicki and Stephen Chong. 2013. [Asynchronous functional reactive programming for guis](#). *SIGPLAN Not.*, 48(6):411–422.
- Nicholas Deas, Jessi Grieser, Xinmeng Hou, Shana Kleiner, Tajh Martin, Sreya Nandanampati, Desmond Patton, and Kathleen McKeown. 2024. [Phonate: Impact of type-written phonological features of african american language on generative language modeling tasks](#). In *Proceedings of the First Conference on Language Modeling*.
- Nicholas Deas, Jessica Grieser, Shana Kleiner, Desmond Patton, Elsbeth Turcan, and Kathleen McKeown. 2023. [Evaluation of African American language bias in natural language generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6805–6824, Singapore. Association for Computational Linguistics.
- Nicholas Deas, Blake Vente, Amith Ananthram, Jessica A Grieser, Desmond U. Patton, Shana Kleiner, James R. Shepard III, and Kathleen McKeown. 2025. [Data caricatures: On the representation of African American language in pretraining corpora](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29192–29217, Vienna, Austria. Association for Computational Linguistics.
- Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. [GoEmotions: A dataset of fine-grained emotions](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.
- Yanai Elazar, Akshita Bhagia, Ian Helgi Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Evan Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, et al. 2024. [What’s in my big data?](#) In *The Twelfth International Conference on Learning Representations*.
- Ben Eyal, Moran Mahabi, Ophir Haroche, Amir Bachar, and Michael Elhadad. 2023. [Semantic decomposition of question and SQL for text-to-SQL parsing](#). In *Findings of the Association for Computational*

- Linguistics: EMNLP 2023*, pages 13629–13645, Singapore. Association for Computational Linguistics.
- Charlie Farrington and Tyler Kendall. 2021. [The corpus of regional african american language](#).
- Simon Fowler. 2020. Model-view-update-communicate: Session types meet the elm architecture (artifact). Schloss Dagstuhl-Leibniz-Zentrum für Informatik.
- Ge Gao, Alexey Taymanov, Eduardo Salinas, Paul Mineiro, and Dipendra Misra. 2024. Aligning llm agents by learning latent preference from user edits. In *Conference on Neural Information Processing Systems*.
- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. [Datasheets for datasets](#). *Commun. ACM*, 64(12):86–92.
- Jessica A Grieser, James Shepard, Nicholas Deas, Shana Kleiner, Desmond Patton, Elsbeth Turcan, and Kathleen McKeown. 2023. Exploring the deficiencies of large language models in use of aal: An interdisciplinary study. In *New Ways of Analyzing Variation (NWAV51)*, Queens College of the City University of New York.
- Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. 2024. [OLMo: Accelerating the science of language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand. Association for Computational Linguistics.
- Sirui Hong, Yizhang Lin, Bang Liu, Bangbang Liu, Binhao Wu, Ceyao Zhang, Danyang Li, Jiaqi Chen, Jiayi Zhang, Jinlin Wang, Li Zhang, Lingyao Zhang, Min Yang, Mingchen Zhuge, Taicheng Guo, Tuo Zhou, Wei Tao, Robert Tang, Xiangtao Lu, Xiawu Zheng, Xinbing Liang, Yaying Fei, Yuheng Cheng, Yongxin Ni, Zhibin Gou, Zongze Xu, Yuyu Luo, and Chenglin Wu. 2025. [Data interpreter: An LLM agent for data science](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19796–19821, Vienna, Austria. Association for Computational Linguistics.
- Shana Kleiner, Jessica A. Grieser, Shug Miller, James Shepard, Javier Garcia-Perez, Nick Deas, Desmond U. Patton, Elsbeth Turcan, and Kathleen McKeown. 2024. [Unmasking camouflage: exploring the challenges of large language models in deciphering african american language & online performativity](#). *AI and Ethics*, 5(1):29–37.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Haokun Liu, Yangqiaoyu Zhou, Mingxuan Li, Chenfei Yuan, and Chenhao Tan. 2025a. [Literature meets data: A synergistic approach to hypothesis generation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 245–281, Vienna, Austria. Association for Computational Linguistics.
- Jiacheng Liu, Taylor Blanton, Yanai Elazar, Sewon Min, Yen-Sung Chen, Arnavi Chheda-Kothary, Huy Tran, Byron Bischoff, Eric Marsh, Michael Schmitz, Cassidy Trier, Aaron Sarnat, Jenna James, Jon Borchardt, Bailey Kuehl, Evie Yu-Yen Cheng, Karen Farley, Taira Anderson, David Albright, Carissa Schoenick, Luca Soldaini, Dirk Groeneveld, Rock Yuren Pang, Pang Wei Koh, Noah A. Smith, Sophie Lebrecht, Yejin Choi, Hannaneh Hajishirzi, Ali Farhadi, and Jesse Dodge. 2025b. [OLMoTrace: Tracing language model outputs back to trillions of training tokens](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 178–188, Vienna, Austria. Association for Computational Linguistics.
- Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2024. [Infini-gram: Scaling unbounded n-gram language models to a trillion tokens](#). Preprint, arXiv:2401.17377.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach](#). Preprint, arXiv:1907.11692.
- Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. [TimeLMs: Diachronic language models from Twitter](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260, Dublin, Ireland. Association for Computational Linguistics.
- Kathleen McKeown, Jessica A. Grieser, Nicholas Deas, Shana Kleiner, Desmond U. Patton, James Shepard, and Blake Vente. 2026. [How well do large language models understand african american language? causes and implications](#). *Annual Review of Linguistics*, 12(1):273–294.

- Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. [SemEval-2016 task 6: Detecting stance in tweets](#). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.
- Hannes Mühleisen and Mark Raasveldt. 2026. [duckdb: DBI Package for the DuckDB Database Management System](#). R package version 1.5.4.3.
- Desmond Patton, Shana Kleiner, Shug Miller, Nicholas Deas, Jessica Grieser, James Shepard, Blake Vente, and Kathleen McKeown. 2025a. [Ai with heart: A joy-centered framework for ethical ai development & implementation](#). *The Paris Journal on AI & Digital Ethics*.
- Desmond Patton, Shana Kleiner, Shug Miller, Nick Deas, Fahnmusa Edwards, Jessi Grieser, James Shepard, Elsbeth Turcan, and Kathy McKeown. 2024. [Digital narratives of grief and resilience: Insights from the integrating emotional stories online \(ieso\) platform](#).
- Desmond Patton, Shana Kleiner, Shug Miller, Nick Deas, Jessie Grieser, James Shepherd, Elsbeth Turcan, and Kathleen McKeown. 2025b. [Detecting grief online among black harlem residents](#). *Biological Psychiatry*, 97(9):S44.
- Ellie Pavlick and Joel Tetreault. 2016. [An empirical analysis of formality in online communication](#). *Transactions of the Association for Computational Linguistics*, 4:61–74.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *NeurIPS EMC² Workshop*.
- Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. [CARER: Contextualized affect representations for emotion recognition](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.
- Melanie Subbiah, Haaris Mian, Nicholas Deas, Ananya Mayukha, Dan P. McAdams, and Kathleen McKeown. 2026. [Whose story gets told? positionality and bias in llm summaries of life narratives](#). *Preprint*, arXiv:2604.20131.
- Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. [Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions](#). In *Proceedings of the 25th International Conference on World Wide Web, WWW '16*, page 613–624, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.
- Xuemei Tang, Xufeng Duan, and Zhenguang Cai. 2025. [Large language models for automated literature review: An evaluation of reference generation, abstract writing, and review composition](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1602–1617, Suzhou, China. Association for Computational Linguistics.
- Laurens van der Maaten. 2014. [Accelerating t-sne using tree-based algorithms](#). *Journal of Machine Learning Research*, 15(93):3221–3245.
- David van Dijk and Ivan Vrkic. 2026. [Scidesign-bench: Benchmarking and improving language models for scientific inverse design](#). *Preprint*, arXiv:2603.12724.
- Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. [SemEval-2018 task 3: Irony detection in English tweets](#). In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.
- Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. [Learning from the worst: Dynamically generated datasets to improve online hate detection](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics.
- Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018. [RtGender: A corpus for studying differential responses to gender](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, Anima Anandkumar, Karianne Bergen, Carla P. Gomes, Shirley Ho, Pushmeet Kohli, Joan Lasenby, Jure Leskovec, Tie-Yan Liu, Arjun Manrai, Debora Marks, Bharath Ramsundar, Le Song, Jimeng Sun, Jian Tang, Petar Veličković, Max Welling, Linfeng Zhang, Connor W. Coley, Yoshua Bengio, and Marinka Zitnik. 2023. [Scientific discovery in the age of artificial intelligence](#). *Nature*, 620(7972):47–60.
- Weiqi Wang, Jiefu Ou, Yangqiu Song, Benjamin Van Durme, and Daniel Khashabi. 2026. [arXiv2Table: Toward realistic benchmarking and evaluation for LLM-based literature-review table generation](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7602–7624, San Diego, California, United States. Association for Computational Linguistics.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini,

- Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. [Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.
- Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.
- Hao Xu, Jiacheng Liu, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. [Infini-gram mini: Exact n-gram search at the Internet scale with FM-index](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24944–24969, Suzhou, China. Association for Computational Linguistics.
- Changbing Yang, Franklin Ma, Freda Shi, and Jian Zhu. 2025. [LingGym: How far are LLMs from thinking like field linguists?](#) In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1314–1340, Suzhou, China. Association for Computational Linguistics.
- Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. [SemEval-2019 task 6: Identifying and categorizing offensive language in social media \(OffensEval\)](#). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. 2024. [Hypothesis generation with large language models](#). In *Proceedings of EMNLP Workshop of NLP for Science*.

A Original Interface Details

The original structured interface design is shown in Figure 5. Users were provided with three stages to guide their data analyses: 1) preparing a dataset, 2) applying labels and filtering, and 3) generating plots or statistical results.

Figure 5: Screenshot of initial version of CorpusLens.

B High-Level Architectural Details

The essential constraints of the system are as follows:⁷

1. Users and LLM-based agents should have equal ability to take actions in the system, including user interface actions and database actions
2. The entire state of the system should be equally visible to all actors, including client side actions taken by the user
3. Models and Users differ in their preferred modalities: it might be best to show visual

⁷A detailed discussion of these and similar constraints can be seen in Czaplicki and Chong

plots to users and best to show data tables to models

As such, the system’s architecture is inspired by The Elm Architecture (Model-View-Update, as in Fowler). All inputs to the system: User Messages, Model messages, Model Context Protocol Tool calls, Database Query output are unified in a single event log that mediates the information available to the user and the model. This is what allows the model to collapse produce and interact with UI components, and what allows the user’s history of selections to be visible to the model. The net effect is that the system can create names groups of records and add to this virtual set interactively across many different turns.

CorpusLens is model-agnostic, owing to its use of the conventionally used LiteLLM library. CorpusLens also uses the official libraries for the Model Context Protocol, to provide COMPASS all of the tools it uses to interact with the system.

C Tool Details

A variety of tools are available to COMPASS in order to interpret, process, and respond to user queries. Namely, COMPASS is provided tools for interacting with the interface, documentation of methodologies, and the database, which are listed in Table 2. While CorpusLens is extensible and allows additional analysis tools, we include an initial set of analysis tools including emotion/sentiment classifiers (Demszky et al., 2020; Saravia et al., 2018; Barbieri et al., 2020; Loureiro et al., 2022), semantics and style features (Zampieri et al., 2019; Basile et al., 2019; Vidgen et al., 2021; Van Hee et al., 2018; Borkan et al., 2019; Mohammad et al., 2016; Antypas et al., 2022; Williams et al., 2018; Pavlick and Tetreault, 2016), dialect classifiers (Blodgett et al., 2016), and embedding reduction techniques (van der Maaten, 2014; Campello et al., 2013) using a variety of pretrained models (Liu et al., 2019; Warner et al., 2025; Sanh et al., 2019; Lewis et al., 2020).

D Human Evaluation Survey

All questions asked of domain experts after using CorpusLens are listed below. The survey includes six Likert-scale questions as well as two open-ended questions for additional comments and feedback.

Category	Technical Reference	Title	Description
UI Operations	<code>collapse_result</code>	Tidying Up	Folds a superseded call into a one-liner.
Corpus Documentation	<code>corpus_stats</code>	Counting the Corpus	Row count and classifier-column count.
	<code>get_schema</code>	Checking the Catalog	Inspects the corpus tables and columns.
Transparency	<code>show_work</code>	Showing the Work	Recalls a past tool call as a labeled card.
	<code>plan</code>	Making a Plan	Records a one-paragraph plan for this turn.
Provenance/ Capabilities	<code>get_provenance</code>	Tracing Provenance	Maps classifier columns to source papers.
	<code>get_recipes</code>	Listing Chart Recipes	Lists ready-to-render chart recipes.
	<code>show_methods</code>	Showing Methods	Charts the research papers behind the corpus.
User Selections	<code>explore_embeddings</code>	Mapping the Embedding Space	Brushable scatter map of the corpus.
	<code>get_selection</code>	Recalling a Selection	Recalls a stored selection by id.
	<code>record_brush_selection</code>	Recording a Selection	Saves a brush region as a named selection.
	<code>list_selections</code>	Listing Selections	Lists stored selections in this session.
Charts & Data Viewing	<code>render_chart</code>	Show Chart	Runs a SQL query and shows the rows as a chart.
	<code>show_chart</code>	Showing Saved Chart	Loads a pre-baked chart by name.
	<code>show_parallel_coords</code>	Comparing Across Axes	Compares classifier scores side by side.
	<code>explore_scores_static</code>	Showing Score Panels	Panel of classifier-score charts.
	<code>inspect</code>	Inspect	Show a record in detail
Database Queries/ Operations	<code>corpus_sample</code>	Sampling the Corpus	Random sample of rows + flagship scores.
	<code>search</code>	Searching the Text	Full-text search across the corpus.
	<code>more_like_this</code>	Finding Similar Items	Hybrid k-nearest by text + embedding.
	<code>query_scores</code>	Checking the database	Asking the corpus a question about existing scores.
	<code>run_sql</code>	Checking the database	Asking the corpus a question.

Table 2: Details of select agent tools available to COMPASS in CorpusLens. These include tools for retrieving and presenting methodology details, processing user selections in visualizations, and displaying charts among others.

- How would you rate your overall experience using the tool?** ☐ Very poor ☐ Poor ☐ Fair ☐ Good ☐ Very good
- Considering a potential alternative tool that you use most often from the above question, how helpful did you find this tool?** ☐ Much less helpful ☐ Slightly less helpful ☐ About the same ☐ Slightly more helpful ☐ Much more helpful
- Thinking about what you originally set out to do, how intuitive did you find using the tool?** ☐ Very unintuitive ☐ Somewhat unintuitive ☐ Neither unintuitive nor intuitive ☐ Somewhat intuitive ☐ Very intuitive
- How easy did you find it to complete your intended task with the tool?** ☐ Very difficult ☐ Somewhat difficult ☐ Neither easy nor difficult ☐ Somewhat easy ☐ Very easy
- How sufficiently do you feel the interface/-model explained what methods were being used for each analysis?** ☐ Methods were rarely explained ☐ Methods were mostly left unexplained ☐ Some methods were explained, some unexplained ☐ Methods were mostly explained ☐ Methods were thoroughly explained
- How helpful was the tool in allowing you to discover new themes, features, or patterns in the analyzed data?** ☐ Very unhelpful ☐ Unhelpful ☐ Neither helpful nor unhelpful ☐ Helpful ☐ Very helpful
- What insights (themes, patterns, etc.), if any, did you gain from using this tool?**
- (Optional) Do you have any other comments or feedback about your experience with the tool?**