APPLING
International Applied Linguistics Research Conference

https://doi.org/10.29140/9781763711686-8

Revisiting Discourse Analysis for AI Literacy and Ethics: An Emerging Framework for Evaluating Generative AI Output

Christopher Leslie
Prince of Songkla University, Thailand

Abstract

As generative AI (genAI) becomes a routine source of communication support, applied linguistics has an opportunity to build on existing methods to examine the discourse and norms the tools model. This paper proposes prompt-based discourse elicitation (PBDE) as an analytical framework that synthesizes emerging insights into studying genAI output as discourse. Three key practices have begun to emerge from recently published papers: bespoke corpora elicited from standardized prompts, multiple tools to avoid treating one system's output as representative of genAI discourse more broadly, and transparent analytic schemes akin to thematic analysis. Taken together, these practices suggest PBDE as a reusable framework for evaluating genAI discourse. Only a small set of studies combines all three practices, so synthesizing them represents an opportunity to fill a methodological gap. The author has initiated six PBDE studies, combining the key methodological practices to evaluate the discourse produced by genAI. The present paper examines the methodological lessons learned from these projects; the individual studies are being reported separately. The six studies demonstrate how genAI presents itself as a universal interface while quietly standardizing discourse, reflecting technocratic worldviews, constraining forms of professional English, and reducing argumentative diversity. Outputs were also unevenly moderated by safety guardrails, particularly in relation to gender identity. Findings from PBDE studies can expand the scope of genAI literacy. In addition to warning students about potential errors or hallucinations, educators can show how genAI output can reflect limited worldviews, constrained English, and stance-

E-mail: christopher.l@psu.ac.th

In Y. Wang, J. Matthews, & B. Zou (Eds.), *Proceedings of the 2nd International Applied Linguistics Research Conference, 2026* (pp. 55–63). Castledown.

taking limits. This awareness can help students preserve and enhance their rhetorical agency.

Keywords: AI literacy, discourse analysis, English language learning, generative AI, prompt-based discourse elicitation (PBDE).

Studying GenAI Output as Discourse

This paper considers emerging practices for studying generative AI (genAI) output under the name *prompt-based discourse elicitation* (PBDE). PBDE uses controlled prompts as elicitation instruments to construct comparable corpora of genAI output for interpretation, drawing from applied linguistics, technology studies, and related fields. It is a proposed methodological synthesis that adapts existing practices from pragmatic elicitation, black-box auditing, corpus construction, and qualitative discourse analysis to the study of output. It can study discourse at the textual level (hedging, framing) and at the ideological level (worldviews, identities). PBDE treats genAI output not simply as facts to be checked, but as discourse to be explored and interpreted.

PBDE can be situated within the tradition of black-box auditing. Sandvig et al. (2014) argued that audit methods developed to detect discrimination in housing and employment can be adapted to investigate online systems; Bandy (2021) similarly described algorithmic auditing of automated systems. These approaches, often conducted in fields like information science, are intellectual ancestors of research seeking to evaluate potentially problematic system behaviours through structured elicitation. Whereas compliance-oriented audits evaluate practices against established standards, PBDE uses structured queries to explore discourse patterns without presupposing what those patterns will be. The PBDE studies reported here adapt two audit modes: single queries and simulated-user interactions (also called *sock puppets*). These correspond to the modes in Table 1.

Small-scale probing of black-box systems has produced influential critiques of online discourse. Pariser (2011) and Noble (2018) used human-directed inquiries of search and recommendation services to show that internet tools are not neutral portals. Rogers (2023) probed tools directly with queries that had previously shown biases, while Morrison (2024) suggested that discourse analysis become a part of search-engine evaluation. Whereas systems audits involve numerous automated queries, this research demonstrates the effectiveness of purposeful, small-scale probing. Like these studies, PBDE does not claim to reveal internal workings or to make statistically generalizable findings. Instead, it uses structured queries to explore genAI discourse and identify patterns that emerge from the resulting outputs.

Strategies for PBDE

Exemplars of recently published studies were identified through targeted reading and were selected because they demonstrate one or more practices relevant to PBDE.

Bespoke Corpora from Standardized Prompts

The first key aspect of PBDE is creating a corpus from responses to fixed prompts. Here, corpus means a deliberately elicited set of outputs, not a statistically representative quantity. One exemplar demonstrated "epistemic oppression" (Carbajal-Carrera, 2024, p. 360) by using a career-advice prompt to compare how genAI valued linguistic identities. Other studies examined fictionalized letters of recommendation, news articles, and patient profiles (Fang et al., 2024; Mao et al., 2026; Putland et al., 2025). Prompt-sensitivity research shows that minor wording changes can affect outputs (e.g., Mizrahi et al., 2024). PBDE makes a different claim: standardized prompts create fixed elicitation conditions for comparing discourse.

Multiple Tools

Several researchers have compared outputs from more than one tool, an important second aspect of PBDE. Exemplary studies show how cross-tool probing can reveal biases and differences in output forms (Dwivedi et al., 2023; Facchinetti, 2025; Lang et al., 2023; Lysova et al., 2026). For example, exploring two different tools helped authors notice bias in gender representation (Fatima et al., 2025). Probing multiple tools denaturalizes genAI discourse, exposing variations that a single-tool study might overlook.

Transparent Analytic Schemes

Because genAI tools change rapidly, PBDE depends on transparent analytic schemes, a third important aspect of PBDE. Rubrics, codebooks, and coded examples make the analysis reusable even when exact outputs cannot be regenerated (Rogers, 2023; Wilson & Rose, 2025). Repeatability is a known problem in black-box auditing, but this issue can be mitigated by using rubrics with clear coding definitions.

Studies Combining All Components

A few researchers have used all three of the basic components. Ting et al. (2025) studied the output of two tools, then conducted inductive thematic analysis, creating a codebook for feedback on student writing. Chatzikoumi and Gonzalez Arias (2026) used controlled interactions in Spanish to produce

a corpus from two tools, which they analysed with a rubric to compare the prevalence of features. Giacalone (2026) used standard prompts on five different tools, evaluating with the assistance of a rubric to show how tone and framing shape meaning.

PBDE Workflow

Prompt-based discourse elicitation adapts familiar applied-linguistic practices used in interviews, surveys, corpus construction, and thematic analysis. Like discourse-completion tasks in cross-cultural pragmatics (Blum-Kulka & Olshtain, 1984), PBDE uses controlled situations to elicit comparable language data, but it studies genAI-generated text rather than human speech acts. As shown in Figure 1, PBDE begins with the design of standardized prompts and a draft coding rubric, both derived deductively from the research questions.

The prompts are presented to multiple genAI tools, generating a corpus of comparable outputs. The rubric guides analysis, with categories refined and elaborated inductively as patterns emerge from the data. This coding becomes discourse analysis when categories attend not only to content, but also to how outputs frame situations, position users, represent identities, assign agency, and construct norms of appropriate language or action. This process supports auditable interpretation because prompts, outputs, coding categories, revisions, and examples are visible to readers. Quantitative analysis can then be supplemented with qualitative interpretation.

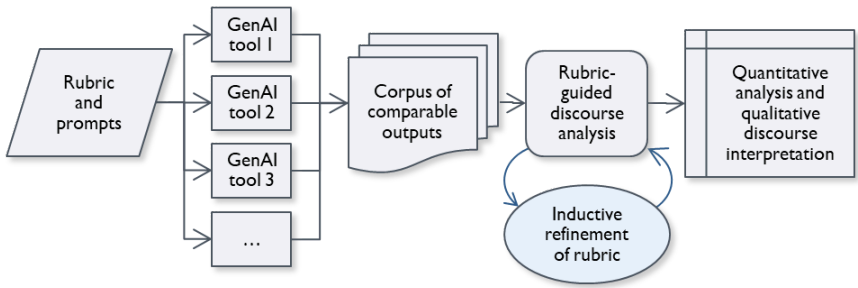

Figure 1. PBDE Workflow

Illustrative Applications

Table 1 summarizes six PBDE studies the author initiated in 2026 as illustrative case studies demonstrating the framework's application.

Table 1. PBDE Studies

Prompt focus	***Discourse concerns***	***Elicitation mode***
1. Task-based learning	Locality and globalized discourse	Single queries
2. Intercultural communication	Neutrality, politeness, conflict management	
3. Choosing a major	Neoliberal frames, localization, user agency	
4. Revising articles	Voice, stance, standardization	
5. Framing of innovation	Progress narratives, Whig history framing	Simulated-user inquiries
6. Helping a student with writing	Guardrails, responsible use	

One recurrent finding is that the apparent universality of genAI tools does not hold up in the context of corpora built from multiple tools in different locations. Studies 1 and 3 showed unevenness in responses tied to specific locations, suggesting the influence of an uneven information landscape previously seen in global media. Importantly, the corpus in Study 3 showed a concerning lack of guardrails with regard to an LGBTQ+ individual in Brunei, telling a prospective student to seek out collegiate and professional organizations in a country where same-sex identities are criminalized.

The studies have led to opportunities for pedagogical innovation based on analyses of the corpora. Study 2 showed that prompts related to intercultural competency could introduce students to unfamiliar viewpoints. A content analysis might code these outputs by topic, such as "politeness", "conflict avoidance", or "directness". A discourse analysis instead asks how the outputs frame intercultural conflict, what kinds of agency they assign to speakers, and whether they model disagreement as negotiation, accommodation, or confrontation. Study 4 showed the homogenizing effect on scholarly writing. Study 6 suggested important interventions that can be used in writing classrooms.

A limitation of the first four studies is their single-shot approach, with one prompt and its response consisting of a unit of analysis. Although this aids coding and interpretation, it does not reflect the way that many users interact with genAI tools. Accordingly, Studies 5 and 6 have structured follow-up questions to see what happens when tools are challenged.

Lessons Learned

PBDE seeks to produce discourse analysis of output that reflects auditable analysis: interpretations are made traceable through standardized prompts, the resulting corpora, and documented analytic decisions.

1. **Number of Coders.** PBDE-type studies vary in coding practices: some use single coders (Facchinetti, 2025; Giacalone, 2026; Rogers, 2023), some multiple coders without intercoder reliability measurements (Chatzikoumi & Gonzalez Arias, 2026; Wilson & Rose, 2025), and others report intercoder reliability (Lang et al., 2023; Lee et al., 2025; Ting et al., 2025). This variation does not indicate that some studies lack rigour. PBDE seeks nuances in output that reveal recurring discourse patterns, so it is not an instrument-validation project. If multiple coders are desired, a negotiated agreement or consensus approach can be used (Campbell et al., 2013; O'Connor & Joffe, 2020). Intercoder reliability shows the extent to which coders agree; PBDE seeks auditable analysis by making interpretations checkable.
2. **Replicability.** Repeatability is a known limitation of black-box audits (see Bandy, 2021, p. 74:23) because vendors tune their products regularly. PBDE therefore seeks to take a snapshot of potentially vexing behaviours. A study may not be completely replicable, but it can be verifiable if researchers provide the chatbot transcripts and report the elicitation record: exact prompts, tool or interface, model version when available, collection date, and any default settings that were changed. Generation parameters such as temperature, top-p, and seeds are relevant to large-scale testing of models through application programming interfaces (APIs), but they are generally not configurable by users of consumer interfaces. The inability to report these parameters in consumer-interface PBDE studies should not be considered a methodological weakness.
3. **Human Scale.** Small-scale investigations are opportunities for inductive exploration more than tests of predetermined assump-

tions. A contained corpus permits close reading while making coding and interpretation transparent. Researchers in information science seek broad tendencies at larger scales, but PBDE pursues nuances.

4. **Pedagogy.** Recurring discourse patterns translate into evidence-informed pedagogical questions and practices. In this sense, PBDE is not only a research framework but also a foundation for pedagogy. By making recurrent patterns in output visible, it helps educators move beyond warnings about hallucination and plagiarism towards classroom attention to discourse, stance, communicative norms, and rhetorical agency.

PBDE therefore offers applied linguistics researchers a practical way to connect AI-output analysis with discourse studies. GenAI literacy requires more than recognizing when the facts are wrong; it also requires recognizing the discourses that chatbots normalize.

References

Bandy, J. (2021). Problematic machine behavior: A systematic literature review of algorithm audits. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), Article 74. https://doi.org/10.1145/3449148

Blum-Kulka, S., & Olshtain, E. (1984). Requests and apologies: A cross-cultural study of speech act realization patterns (CCSARP). *Applied Linguistics, 5*(3), 196–213. https://doi.org/10.1093/applin/5.3.196

Campbell, J. L., Quincy, C., Osserman, J., & Pedersen, O. K. (2013). Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research, 42*(3), 294–320. https://doi.org/10.1177/0049124113500475

Carbajal-Carrera, B. (2024). AIsplaining: Generative AI explains linguistic identities to me. *Australian Review of Applied Linguistics, 47*(3), 340–365. https://doi.org/10.1075/aral.24077.car

Chatzikoumi, E., & Gonzalez Arias, C. A. (2026). Limits of the anthropomorphisation of discourse in chatbots. In B. Karamifar & A. C. Valente (Eds.), *Artificial intelligence and discourse: Volume 2, Critical views of AI integrated communication* (pp. 127–157). Palgrave Macmillan. https://doi.org/10.1007/978-3-032-07761-5_5

Dwivedi, S., Ghosh, S., & Dwivedi, S. (2023). Breaking the bias: Gender fairness in LLMs using prompt engineering and in-context learning.

Rupkatha Journal on Interdisciplinary Studies in Humanities, 15(4). https://doi.org/10.21659/rupkatha.v15n4.10

Facchinetti, R. (2025). Interviewing AI: A new challenge for discourse analysis in the neural age. *International Journal of Language Studies, 19*(2), 13–30. https://doi.org/10.5281/zenodo.15250779

Fang, X., Che, S., Mao, M., Zhang, H., Zhao, M., & Zhao, X. (2024). Bias of AI-generated content: An examination of news produced by large language models. *Scientific Reports, 14*(1)*,* Article 5224. https://doi.org/10.1038/s41598-024-55686-2

Fatima, W., Yasmin, M., & Irshad, I. (2025). Exploring gendered language and socio-cognitive impact in AI-generated texts: A critical discourse approach. *Pakistan Languages and Humanities Review, 9*(3), 504–524. https://doi.org/10.47205/plhr.2025(9-III)42

Giacalone, M. (2026). Discursive behavior of generative language models in geopolitical and humanitarian contexts. *Discover Artificial Intelligence, 6,* Article 230. https://doi.org/10.1007/s44163-026-00965-2

Lang, G., Triantoro, T., & Sharp, J. H. (2023). A comparative analysis of large language models for writing teaching cases. *Proceedings of the ISCAP Conference, 9*(5948), 1–24. https://iscap.us/proceedings/2023/pdf/5948.pdf

Lee, N. S., Richards, N., Grandominico, J., Cronin, R. M., Hendricks, A. K., Tripathi, R. S., & Jonas, D. E. (2025). Use of a medical communication framework to assess the quality of generative artificial intelligence replies to primary care patient portal messages: Content analysis. *JMIR Formative Research, 9,* Article e71966. https://doi.org/10.2196/71966

Lysova, I., Ahmed, L., Cunningham, B., Huang, Y., & Wiltschko, M. (2026). Do chatbots dream of AI sheep? A semantic–pragmatic investigation of "naturalness" in human–AI interaction. *AI & Society, 41*(2), 1053–1062. https://doi.org/10.1007/s00146-025-02595-1

Mao, E., Glenn, E. R., LaPorte, D. M., Hsu, N. N., Thompson, J. M., & Aiyer, A. (2026). Does ChatGPT show gender bias when drafting letters of recommendation for applicants to orthopaedic surgery residency programs? *JBJS Open Access, 11*(1), Article e25.00319. https://doi.org/10.2106/JBJS.OA.25.00319

Mizrahi, M., Kaplan, G., Malkin, D., Dror, R., Shahaf, D., & Stanovsky, G. (2024). State of what art? A call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics, 12,* 933–949. https://doi.org/10.1162/tacl_a_00681

Morrison, R. (2024). Making the invisible visible: Critical discourse analysis as a tool for search engine research. *Journal of the Association for*

Information Science and Technology, 75(5), 600–612. https://doi.org/10.1002/asi.24859

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.

O'Connor, C., & Joffe, H. (2020). Intercoder reliability in qualitative research: Debates and practical guidelines. *International Journal of Qualitative Methods, 19*, Article 1609406919899220. https://doi.org/10.1177/1609406919899220

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin Press.

Putland, E., Chikodzore-Paterson, C., & Brookes, G. (2025). "We no longer recognized her as a human being": A critical discourse analysis of AI-generated character descriptions of men and women with dementia. *Journal of Language and Discrimination, 9*(2), 183–208. https://doi.org/10.3138/jld-2025-1002

Rogers, R. (2023). Algorithmic probing: Prompting offensive Google results and their moderation. *Big Data & Society, 10*(1). https://doi.org/10.1177/20539517231176228

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014, May 22). *Auditing algorithms: Research methods for detecting discrimination on internet platforms* [Paper presentation]. Data and Discrimination: Converting Critical Concerns into Productive Inquiry, Seattle, WA, United States.

Ting, L., Singh, C. K. S., Kiong, T. T., Rakhimjonov, F., & Singh, T. S. M. (2025). Mentor or examiner? A critical discourse analysis of AI-generated feedback in EFL writing education. *International Journal of Academic Research in Progressive Education and Development, 14*(4), 664–684. https://doi.org/10.6007/IJARPED/v14-i4/26651

Wilson, W., & Rose, H. (2025). A genre, scoring, and authorship analysis of AI-generated and human-written refusal emails. *Business and Professional Communication Quarterly*. Advance online publication. https://doi.org/10.1177/23294906251322890