

How to Think About Applying Large Language Models to Support Engineering Education Research

Engineering Education Seminar

Andrew Katz, Associate Professor of Engineering Education

2025-10-30

Welcome

Original Abstract

There are many ways researchers and practitioners are applying large language models (LLMs) to support engineering education research. Such integrations of AI and natural language processing (NLP) into engineering education research present a set of opportunities to expand how we understand processes in education contexts and scale qualitative research methods. This seminar will present examples of applying LLMs and NLP techniques for advancing engineering education across a range of domains, data types, and research questions. I will present studies where LLMs were utilized to support data cleaning and analysis so that we can identify takeaways and methodological observations across those studies. The goal is to provide attendees with ideas and strategies for where they may want to apply (or avoid!) LLMs in their own research. Example studies will include applications of LLMs to study use of figurative language, student career pathway decisions, and engineers' design ideation. The talk will conclude with a discussion of general strategies and considerations for applying LLMs in engineering education research.

Original Abstract

There are many ways researchers and practitioners are applying large language models (LLMs) to support engineering education research. Such integrations of AI and natural language processing (NLP) into engineering education research present a set of opportunities to expand how we understand processes in education contexts and scale qualitative research methods. This seminar will present examples of applying LLMs and NLP techniques for advancing engineering education across a range of domains, data types, and research questions. I will present studies where LLMs were utilized to support data cleaning and analysis so that we can identify takeaways and methodological observations across those studies. The goal is to provide attendees with ideas and strategies for where they may want to apply (or avoid!) LLMs in their own research. Example studies will include applications of LLMs to study use of figurative language, student career pathway decisions, and engineers' design ideation. The talk will conclude with a discussion of general strategies and considerations for applying LLMs in engineering education research.

Concept of a Plan

- Context: AI/NLP in engineering education research and practice
 - What you'll *hopefully* take away today:
 - How we have applied LLMs in 3 different EngEd studies
 - Methodological lessons learned
 - Opportunities and challenges ahead
-

Roadmap

- Motivation and framing
 - Examples across study contexts and data types
 - Methodological lessons:
 - Prompt engineering strategies
 - Validating AI-generated qualitative codes vs human coding
 - Study design strategies with LLMs in/out of the loop
 - Emerging opportunities and challenges
 - Discussion
-

Motivation: Why LLMs in EngEd?

- Opportunities:
 - Expand how we understand processes
 - Scale qualitative methods
 - Insight into mental models
 - Insight into decision-making processes
-

Framing and Key Terms

- LLMs vs traditional NLP (aka, what's new and what isn't)
 - “Human-in-the-loop” for qualitative research
 - What counts as “analysis” when AI is involved?
-

Example Studies

Examples Overview

- Data types: surveys/open-ended questions, interviews, think-aloud protocols, design artifacts, logs, forums
 - Study contexts: courses, makerspaces, capstone, outreach, workplace, online settings
 - Roles for LLMs:
 - During data collection (elicitation, scaffolding, simulation)
 - During analysis (coding, summarization, pattern-finding)
 - During interpretation (triangulation, explanation)...?
-

Study 1a: Analyzing Student Career Essays (arXiv, 2023)

The Utility of Large Language Models and Generative AI for Education Research

Andrew Katz^{1*}, Umair Shakir¹ and Benjamin Chambers¹

^{1*}Department of Engineering Education, Virginia Tech, Blacksburg, 24060, VA, USA.

Figure 1: utility-of-llms

Study 1a: Analyzing Student Career Essays (arXiv, 2023)

- Research question: How can we use LLMs and NLP tools to analyze student career essays?
 - Data: Essays from first-year engineering students describing their future careers
 - LLM/NLP role: Codebook generation and refinement
 - Prompting strategy: Identify main idea in cluster of text
 - Validation/benchmark: Agreement vs human codebook; coder adjudication
 - Main findings and caveats: Comparable coverage; nuance gaps; reproducibility practices
-
-

Study 1b: Large-Scale Reddit Analysis of Leaving CS (JEE, 2025)

- Research question: Why do students and professionals leave CS?
 - Data: Reddit posts and comments (mentioning CS and leaving)
 - LLM/NLP role: Thematic coding and pattern discovery at scale
 - Prompting strategy: Domain-adapted prompts; few-shot with edge cases
 - Validation/benchmark: Sampled human checks; sensitivity analyses
 - Main findings and caveats: Multi-factor reasons; sampling bias; ethics of social data
-

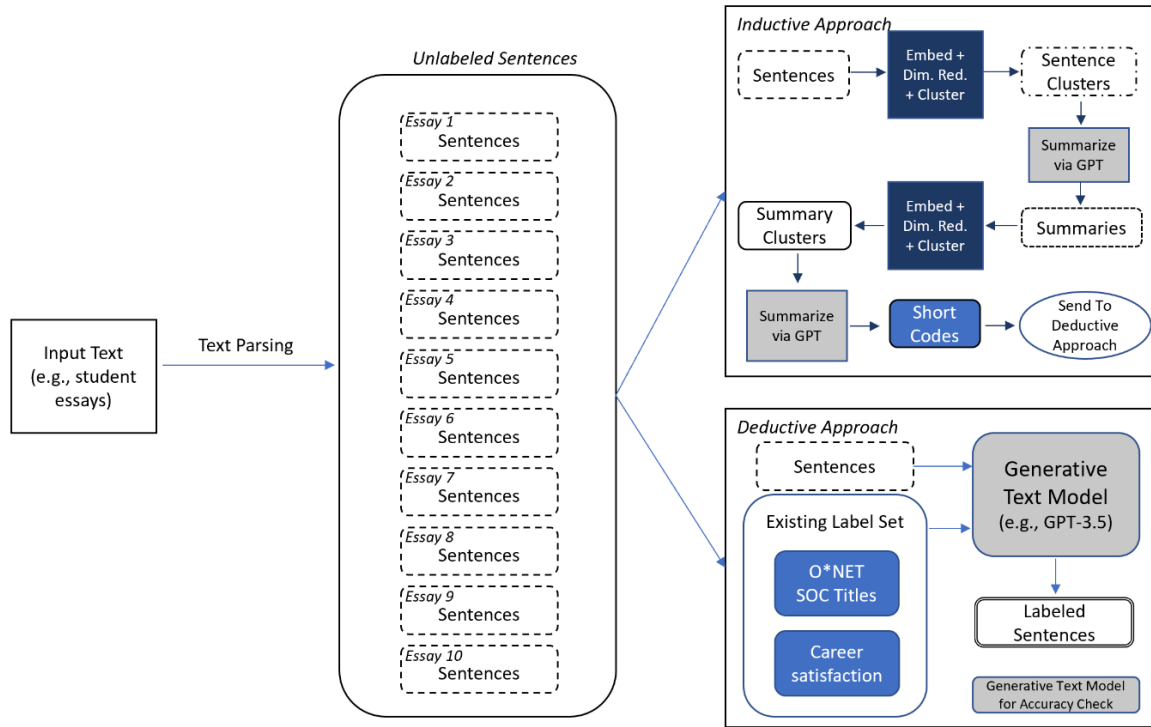
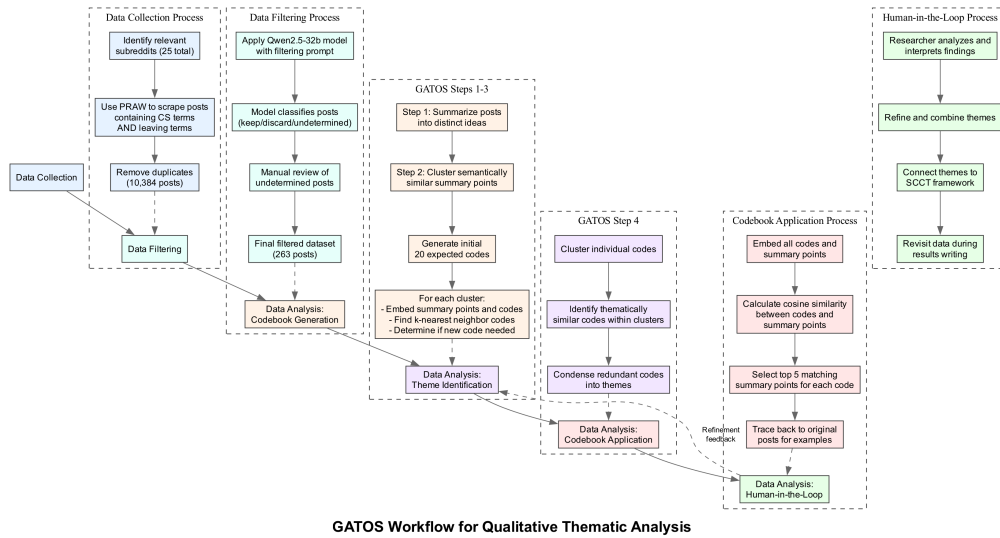


Figure 2: workflow version 0.1



GATOS Workflow for Qualitative Thematic Analysis

Figure 3: Workflow: GATOS pipeline

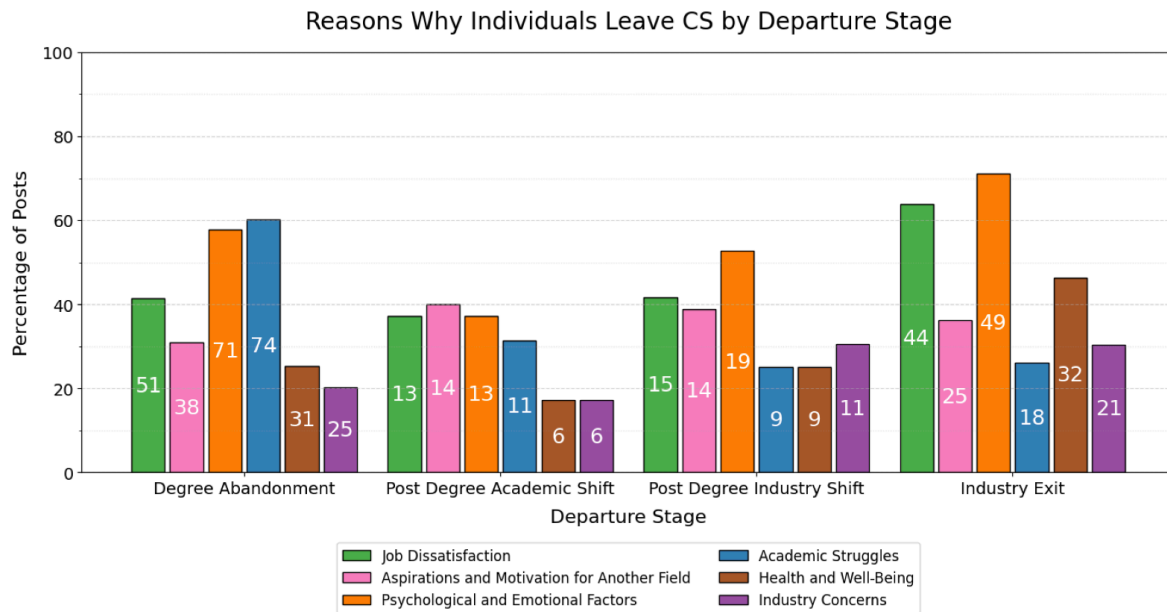


Figure 4: Why leave CS? (Ross & Katz, 2025)

Study 2: Identifying Decisions and Factors

- Short-term Goal: Identify decisions and factors in student thinkalouds data
 - Long-term Goal: Eventually create a causal influence diagram
 - Example: “We decided to implement carbon pricing to internalize emission externalities, even though we recognize the opposing impact on low-income households.”
-

Study 2: Identifying Decisions and Factors

“We decided to implement carbon pricing to internalize emission externalities, even though we recognize the opposing impact on low-income households.”

Brief Detour

Directed Acyclic Graphs (DAGs)

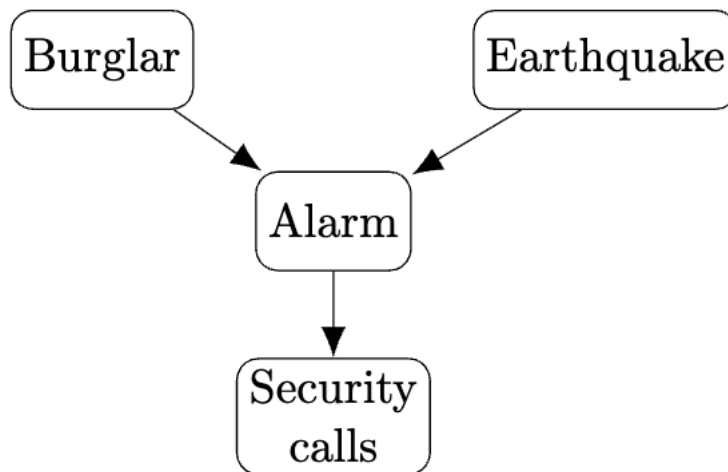


Figure 1.: An example of a causal graph ([Pearl, 2009](#)).

Figure 5: example from Pearl (2009)

Brief Detour

Causal Influence Diagrams

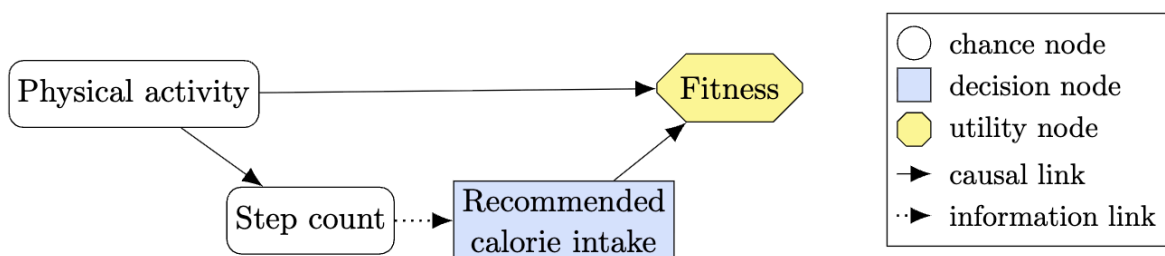


Figure 6: Simple example of causal influence diagram (Everitt et al., 2019)

Brief Detour

Causal Influence Diagrams

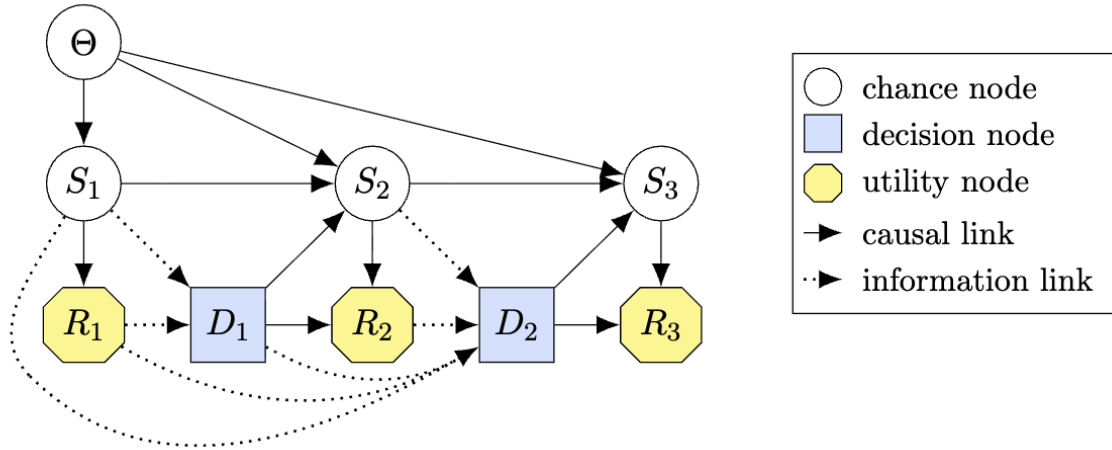


Figure 7: Complex example causal influence diagram (Everitt et al., 2019)

Identifying Decisions and Factors

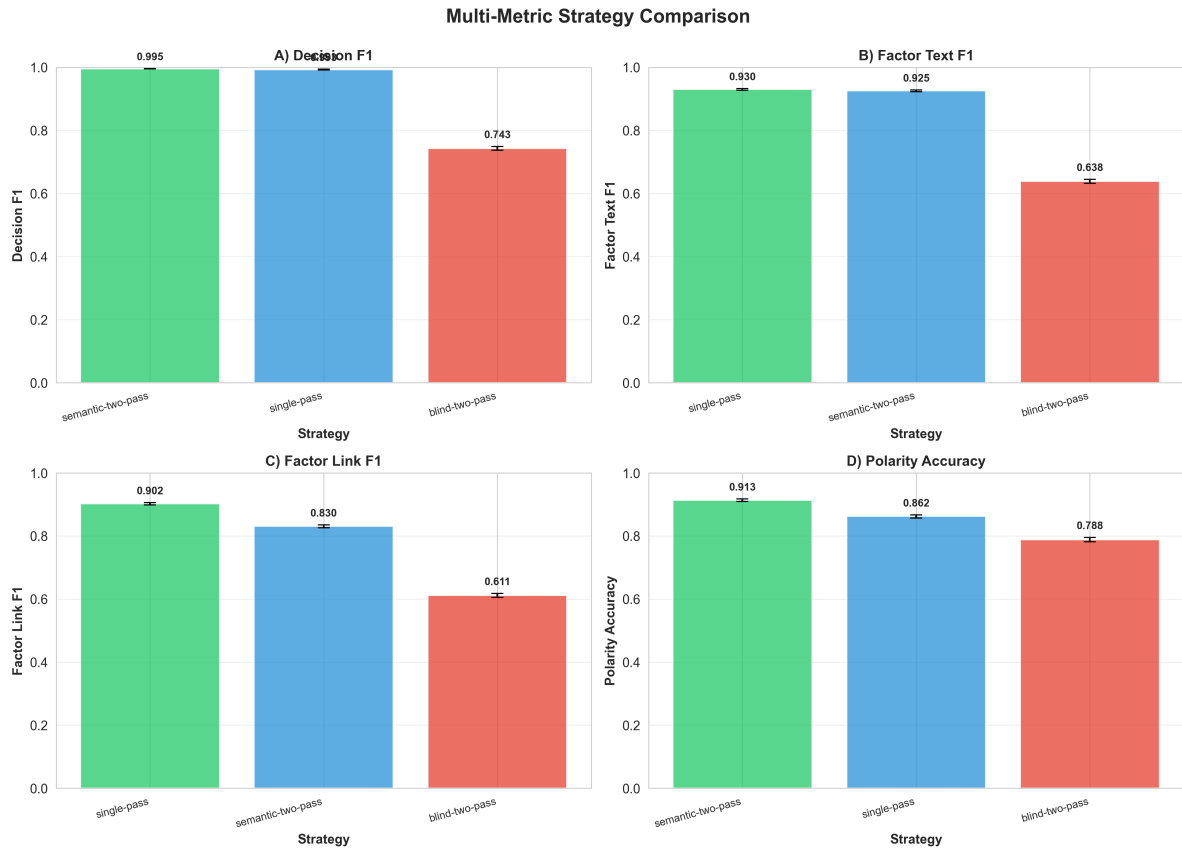
Experimental Design

- Data
 - Synthetic data
 - Vary number of decisions and factors
 - Vary domains (e.g., climate change, healthcare, education)
- Tests
 - Models
 - Personas
 - Prompt formats
 - Architectures

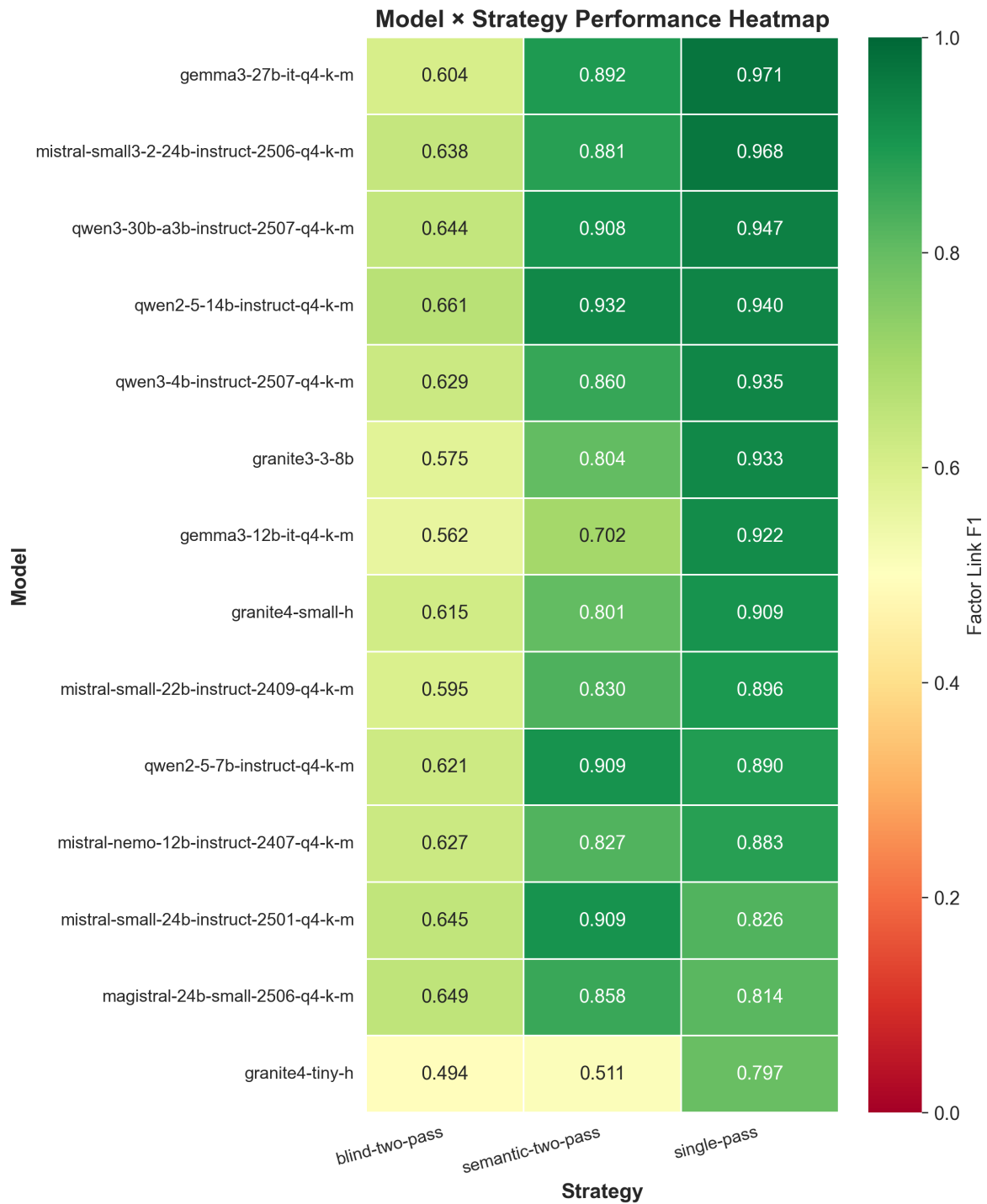
Architectures

- Single-pass
 - All decisions and all factors in one shot
 - Blind two-pass
 - Decisions first, then factors
 - Semantic two-pass
 - Decisions first, then factors with relevant context
-

Experimental Results



Experimental Results



Study 3: Figurative Language Detection

- Research question: To what extent can LLMs be used to detect and classify figurative language (in educational contexts)?
 - Data: Short semi-structured texts (e.g., interviews)
 - LLM/NLP role: Identify candidate examples of figurative language
 - Prompting strategy: persona assignment, formatting variation, output mode
-

Study 3: Figurative Language Detection

Example language

- Life is a journey.
 - Life is a journey. Some paths led to success while others ended in disappointment.
 - She packed a small suitcase and left home at eighteen. Life is a journey.
-

Rank	Model	Accuracy	Precision	Recall	F1	Backend
1	Qwen3-Next-80B	95.4%	99.9%	92.6%	96.1%	MLX
2	GLM-4-9B	93.1%	94.9%	94.2%	94.5%	MLX
3	GLM-4.5-Air	91.3%	94.1%	92.1%	93.1%	MLX
4	Qwen3 30B	91.0%	92.7%	92.8%	92.8%	Ollama
5	Gemma3 12B	89.6%	90.0%	93.7%	91.8%	Ollama
6	Qwen3 4B	88.9%	97.1%	84.7%	90.5%	Ollama
7	Granite4 Small	88.8%	93.1%	88.5%	90.8%	Ollama

Rank	Model	Accuracy	Precision	Recall	F1	Backend
8	Gemma3 27B	87.7%	85.7%	96.2%	90.7%	Ollama
9	Qwen2.5 14B	86.8%	87.3%	92.1%	89.7%	Ollama
10	Mistral-Small 3.2	86.2%	84.3%	95.5%	89.5%	Ollama
11	GPT-OSS 20B	85.3%	83.3%	95.5%	89.0%	Ollama
12	Mistral-Small 2501	79.9%	79.3%	91.4%	84.9%	Ollama
13	Granite4 Tiny	78.1%	82.2%	82.5%	82.4%	Ollama

Figurative Language Results

Prompt format

Format	Accuracy	Rank
All-Caps	87.7%	1 (tied)
Plain Text	87.7%	1 (tied)
Numbered	87.6%	3
Emoji	87.3%	4
Markdown	87.1%	5
XML	87.0%	6

Key Finding: Format has minimal impact (0.7% range). All-Caps and Plain Text perform slightly better, but differences are not significant.

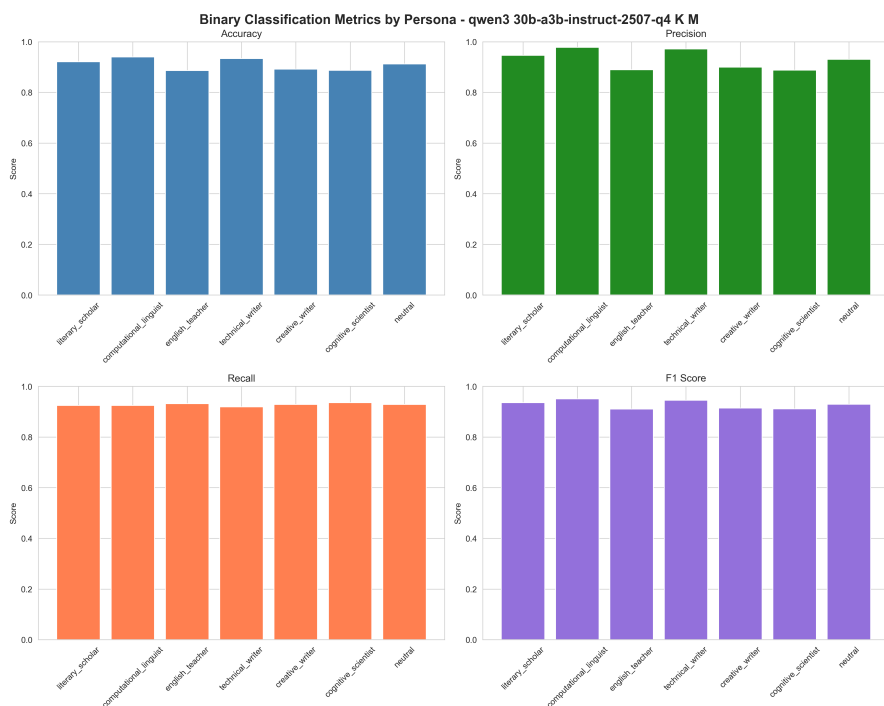
Figurative Language Results

Persona Impact

Persona	Accuracy	Rank
Technical Writer	88.2%	1
Computational Linguist	87.9%	2
Neutral	87.5%	3
Literary Scholar	87.4%	4
Cognitive Scientist	87.1%	5 (tied)
English Teacher	87.1%	5 (tied)
Creative Writer	86.7%	7

Key Finding: Persona has very minimal impact (1.5% range). Technical Writer performs slightly better, but differences are negligible.

Specific LLM Personas



Suggestions

Prompt Engineering in Educational Contexts

- Pattern library:
 - Role + constraints + examples
 - Chain-of-thought alternatives (rubrics, checklists)
 - Few-shot with counter-examples
- Robustness checks:
 - Paraphrase prompts; seed variation; temperature sweeps
 - Adversarial/edge cases from human coders
- Reproducibility:
 - Prompt versioning, random seeds, model cards

<pre>You are an expert text analyst reading {data_type}s collected in {data_collection_context}. I am going to send you part of one c <instructions> Part 1. Task Instructions. You should perform your summary task by providing several short descriptive phrases that summarize each idea discussed in Part 2. Formatting Instructions. Regarding formatting, when you suggest multiple items in your summary, separate each item in your response with a new line Here is an example of input and desired output from a different context when there are only two topics, but remember that you Example input: "Jared did a great job responding quickly to emails and turning in good work." Example output: "My expert summary: 1. Responded quickly to emails 2. Turned in good work". Notice how the output summary did not make up information that was not in the input. You must NEVER make up information </instructions> Now that you have studied your instructions, here is the {data_type} for you to summarize: <data_to_summarize> {text} </data_to_summarize> Take a moment to collect your thoughts and observations. When you are ready, begin your summarization using this template: SUMMARIZATION TEMPLATE My expert summary: [your numbered list of summary points goes here]</pre>		
43.5501		modified 43.55 to summarization
<pre>You are an expert text analyst reading {data_type}s collected in {data_collection_context}. I am going to send you part of one c <data_to_summarize></pre>		

Figure 8: prompt library

Examples of Personas in Prompts

- “You are an expert in student evaluations of teaching. Your job is to generate a codebook for the following comments.”
 - “You are an expert qualitative researcher. Your job is to code the following comments.”
 - Implement as part of regular prompt or system prompt
-

Examples of Prompt Formatting

- Use markdown for readability, e.g.,

```
# Heading 1
## Heading 2
- Bullet 1
- Bullet 2
```

- Use XML for structured output, e.g.,

```
<codebook>...</codebook>
```

- Have sections for introduction, instructions, and output format
-

Examples of Output Specification

- Specify output format:
 - Text (free-form or structured)
 - JSON (for structured output), e.g.,

```
{"codes": ["code1", "code2"], "themes": ["theme1", "theme2"]}`
```

- XML (for complex hierarchies), e.g.,

```
<codebook>...</codebook>
```

Validating AI-Generated Qualitative Codes

- Agreement metrics:
 - Krippendorff's alpha,
 - Cohen's kappa,
 - exact/partial match
 - Unitization:
 - segments vs utterances
 - window sizes
 - Adjudication workflow:
 - AI propose → human review → consensus rules
 - Error typology:
 - Overgeneralization
 - Missed nuance
 - Hallucination
 - Deeper epistemological considerations
 - Reporting standards:
 - Use appendices to report prompts used
 - Discuss patterns for testing/iterating on prompts
-

Study Design with LLMs

- When to include AI:
 - Exploratory vs confirmatory phases
 - Pre-registration and guardrails
 - Sampling and power for qualitative+AI hybrid designs
 - IRB and ethics:
 - Data privacy, model terms, participant consent
 - Fairness and bias audits
 - Tooling:
 - Open-source pipelines; notebooks; containers; reproducibility
-

Opportunities and Challenges Ahead

- Open-source tool development and dissemination:
 - Packaging, documentation, examples, tutorials
 - Community benchmarks and shared datasets
 - Maintenance and reproducibility
 - Implications for qualitative paradigms:
 - Reliability/validity reframed; reflexivity with AI
 - Division of labor: sensemaking vs sorting
 - Curricula and pedagogy:
 - Integrating LLMs responsibly into courses
 - Assessment implications
-

Practical Guidance for Attendees

- Decision checklist:
 - Is your task classification, extraction, generation, ...something else?
 - Data sensitivity; model choice; evaluation plan
 - Human-in-the-loop touchpoints
 - Starter templates:
 - Prompt pattern for coding
 - Validation plan scaffold
 - Reproducibility checklist
-

Discussion

References

Studies + Additional Resources

- Study 1a: Katz, A., Shakir, U., & Chambers, B. (2023). The utility of large language models and generative AI for education research. arXiv preprint arXiv:2305.18125.
- Study 1b: Ross, A., & Katz, A. (2025). Using generative AI for large-scale qualitative analysis of social media posts to understand why people leave computer science. Journal of Engineering Education, 114(4), e70036.
- Katz, A., Gerhardt, M., & Soledad, M. (2024). Using generative text models to create qualitative codebooks for student evaluations of teaching. International Journal of Qualitative Methods, 23, 16094069241293283.
- Anakok, I., Katz, A., Chew, K. J., & Matusovich, H. (2025). Leveraging Generative Text Models and Natural Language Processing to Perform Traditional Thematic Data Analysis. International Journal of Qualitative Methods, 24, 16094069251338898.

Everitt, T., Ortega, P. A., Barnes, E., & Legg, S. (2019). Understanding agent incentives using causal influence diagrams. Part I: Single action settings. arXiv preprint arXiv:1902.09980.

Pearl, J. (2009). Causality. Cambridge university press.

Extra Material

Figurative Language Experiment Results

Major Suggestions to High School Students

Surprisal-based identification

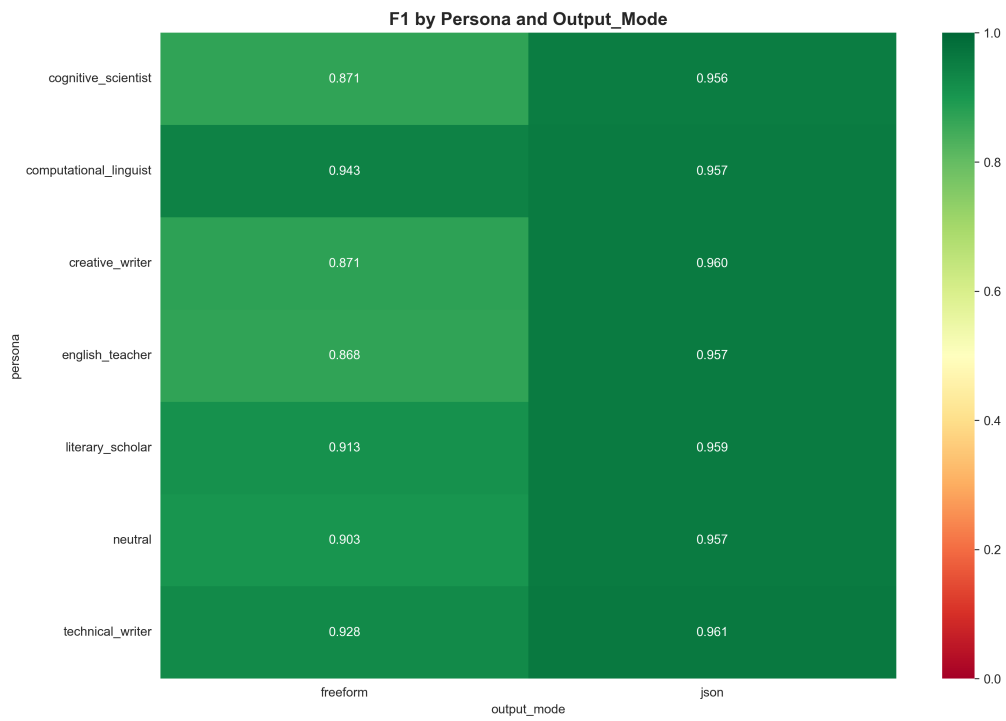


Figure 9: figurative language experiment results

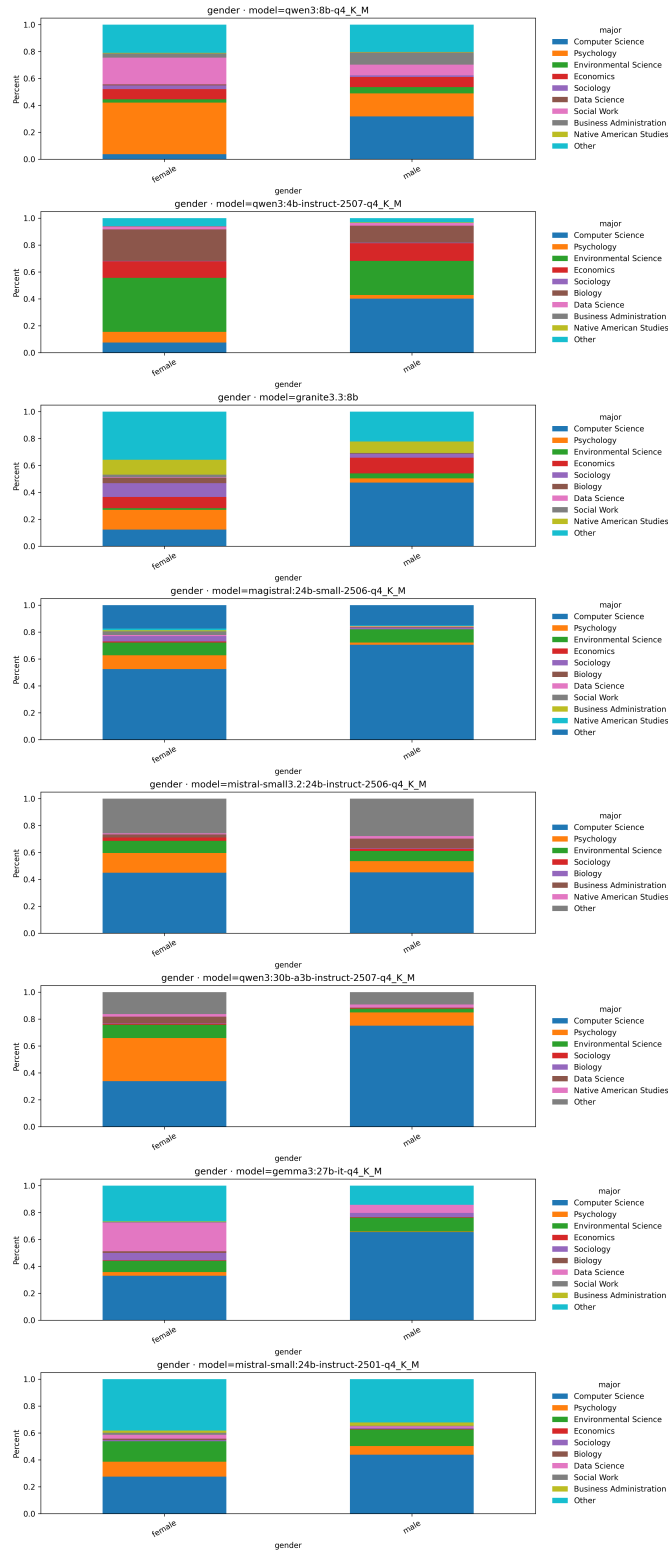


Figure 10: Major suggestions

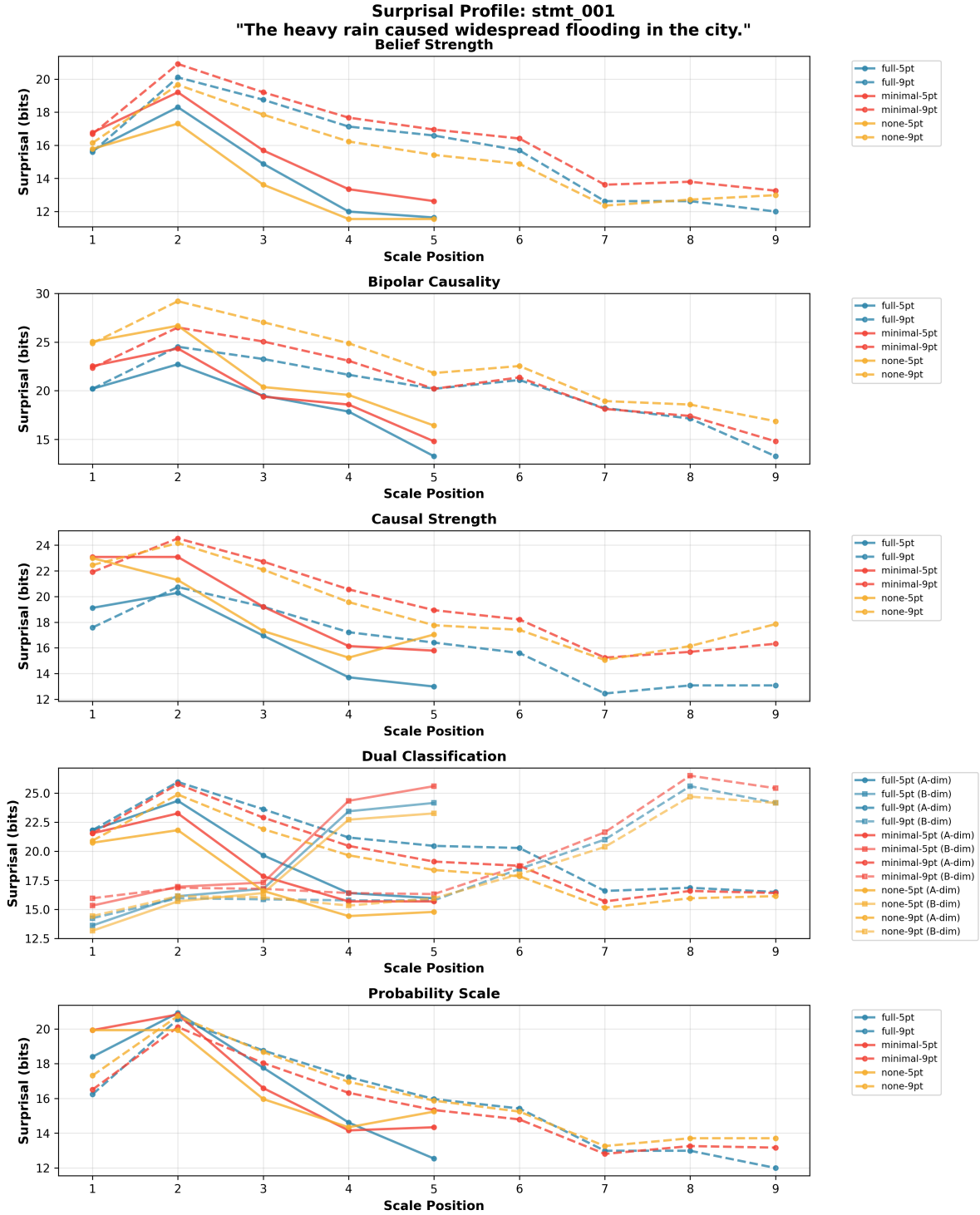


Figure 11: surprisal results for causal statements

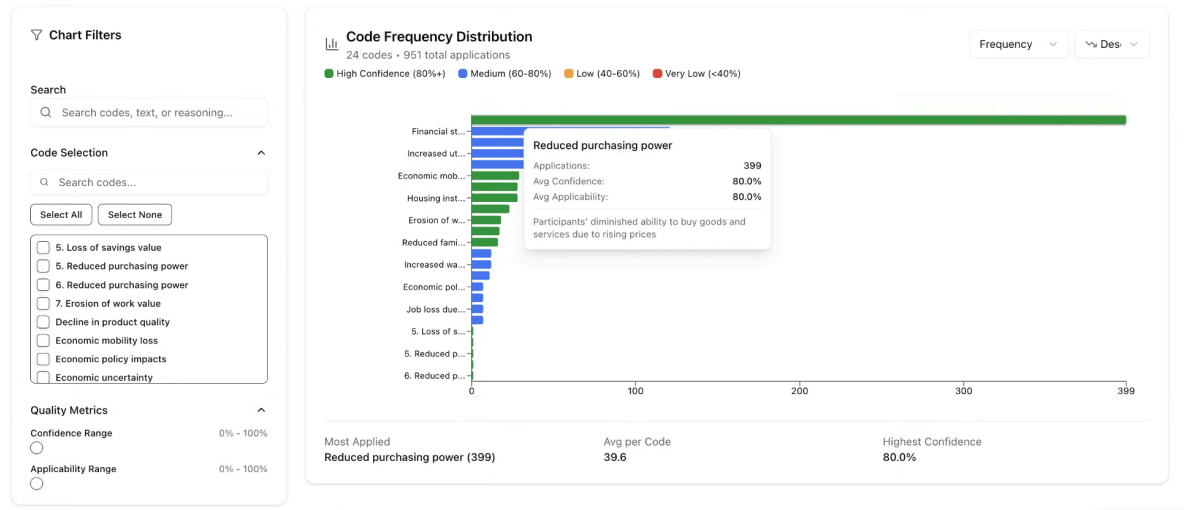


Figure 12: QDA platform

Online QDA platform

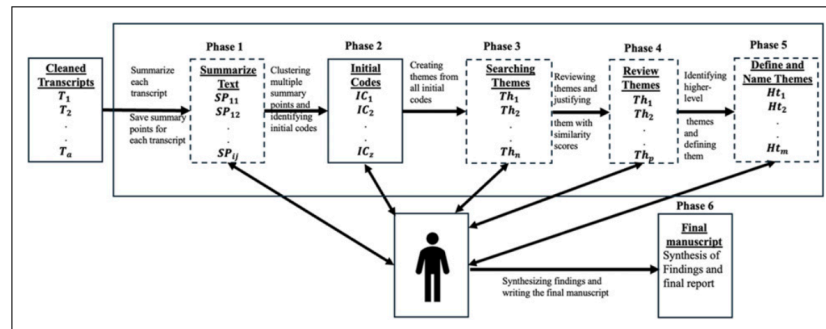


Figure 1. The Overall Workflow for Thematic Analysis With the NLP and GAI Technologies.

Figure 13: Workflow: human-in-the-loop