

From Natural Language to New Understandings: Developing AI-Powered Methodologies for Engineering Education Research and Practice

Andrew Katz, Associate Professor of Engineering Education

2025-11-06

Welcome

Original Abstract

The integration of artificial intelligence and natural language processing (NLP) into engineering education research and practice represents a set of opportunities to expand how we understand student learning, assess educational outcomes, and scale qualitative research methods. This talk presents a comprehensive overview of methodological innovations developed in the IDEEAS Lab over the past five years to leverage the power of generative AI, large language models, and NLP techniques for advancing engineering education. I will present three interconnected methodological contributions: (1) AI-assisted qualitative analysis, where I demonstrate how generative text models can perform traditional thematic data analysis while maintaining rigor and validity; (2) AI-assisted assessment and feedback systems, showcasing applications of open-source language models for analyzing student responses, teamwork feedback, and ethics case studies; and (3) mental models research at scale, illustrating how NLP techniques can help uncover engineering faculty and student mental models of complex topics like sustainability, assessment, and now generative AI itself.

Original Abstract

The integration of artificial intelligence and natural language processing (NLP) into engineering education research and practice represents a set of opportunities to expand how we understand student learning, assess educational outcomes, and scale qualitative research methods. This talk presents a comprehensive overview of methodological innovations developed in the IDEEAS Lab over the past five years to leverage the power of generative AI, large language models, and NLP techniques for advancing engineering education. I will present three interconnected methodological contributions: (1) AI-assisted qualitative analysis, where I demonstrate how generative text models can perform traditional thematic data analysis while maintaining rigor and validity; (2) AI-assisted assessment and feedback systems, showcasing applications of open-source language models for analyzing student responses, teamwork feedback, and ethics case studies; and (3) mental models research at scale, illustrating how NLP techniques can help uncover engineering faculty and student mental models of complex topics like sustainability, assessment, and now generative AI itself.

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 use cases
 - Methodological lessons learned
 - Opportunities and challenges ahead
-

Motivation: Why LLMs in EngEd?

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

Example Use Cases

Examples Overview

- Data types
 - Surveys/open-ended questions
 - Interviews
 - Think-aloud protocols
 - Design artifacts
 - Logs
 - Study contexts
 - Courses
 - Makerspaces
 - Workplace
 - Online settings
 - Roles for LLMs
 - Data collection
 - * Elicitation
 - * Scaffolding
 - * Simulation
 - Analysis
 - * Summarization
 - * Pattern-finding
 - * Coding
-

Use Case 1

“Thematic” Analysis of Large-Scale Qualitative Data

Received: 6 November 2024 | Revised: 25 August 2025 | Accepted: 9 September 2025

DOI: 10.1002/jee.70036

RESEARCH ARTICLE

JEE
Journal of Qualitative Evaluation

Using generative AI for large-scale qualitative analysis of social media posts to understand why people leave computer science

Amanda Ross  | Andrew Katz 

Special Issue: The Evolving Landscape of Qualitative Research: Exploring the Potential and Navigating the Challenges of Artificial Intelligence

Using Generative Text Models to Create Qualitative Codebooks for Student Evaluations of Teaching

International Journal of Qualitative Methods

Volume 23: 1–19

© The Author(s) 2024

DOI: 10.1177/16094069241293283

journals.sagepub.com/home/ijq

 Sage

Andrew Katz¹ , Mitch Gerhardt¹ , and Michelle Soledad¹

Special Issue: The Evolving Landscape of Qualitative Research: Exploring the Potential and Navigating the Challenges of Artificial Intelligence

Leveraging Generative Text Models and Natural Language Processing to Perform Traditional Thematic Data Analysis

International Journal of Qualitative Methods

Volume 24: 1–13

© The Author(s) 2025

DOI: 10.1177/16094069251338898

journals.sagepub.com/home/ijq

 Sage

Isil Anakok¹ , Andrew Katz¹ , Kai Jun Chew², and Holly Matusovich¹

Thematic Analysis with Open-Source Generative AI and Machine Learning: A New Method for Inductive Qualitative Codebook Development

Andrew Katz¹, Gabriella Coloyan Fleming¹, and Joyce Main²

¹Department of Engineering Education, Virginia Tech

²School of Engineering Education, Purdue University

Example Study

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

Use Case 2: Evaluation

Teammate Comment Quality

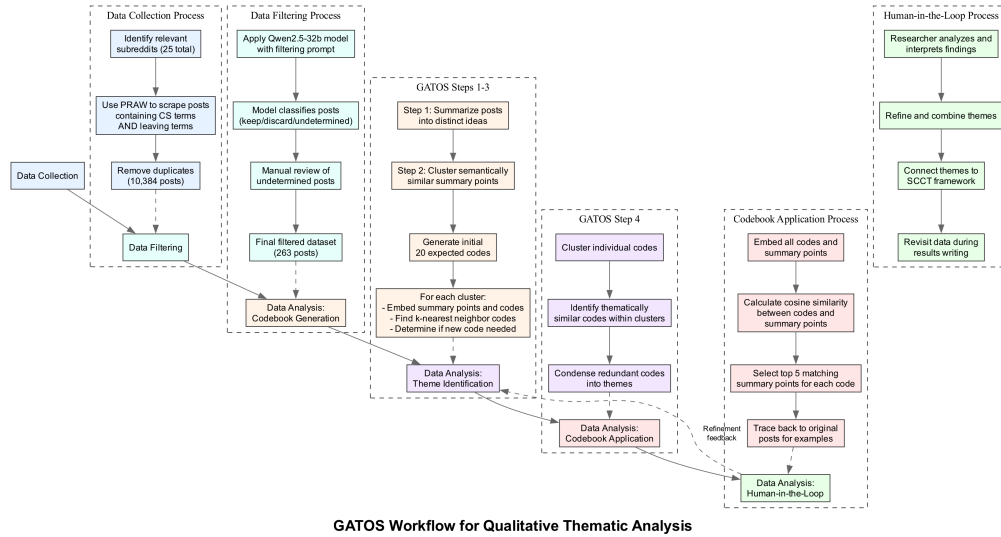


Figure 1: Workflow: GATOS pipeline

Use Case 2: Evaluation

Teammate Comment Quality

Use Case 2: Evaluation

Teammate Comment Quality

Use Case 3:

Mental Models Research

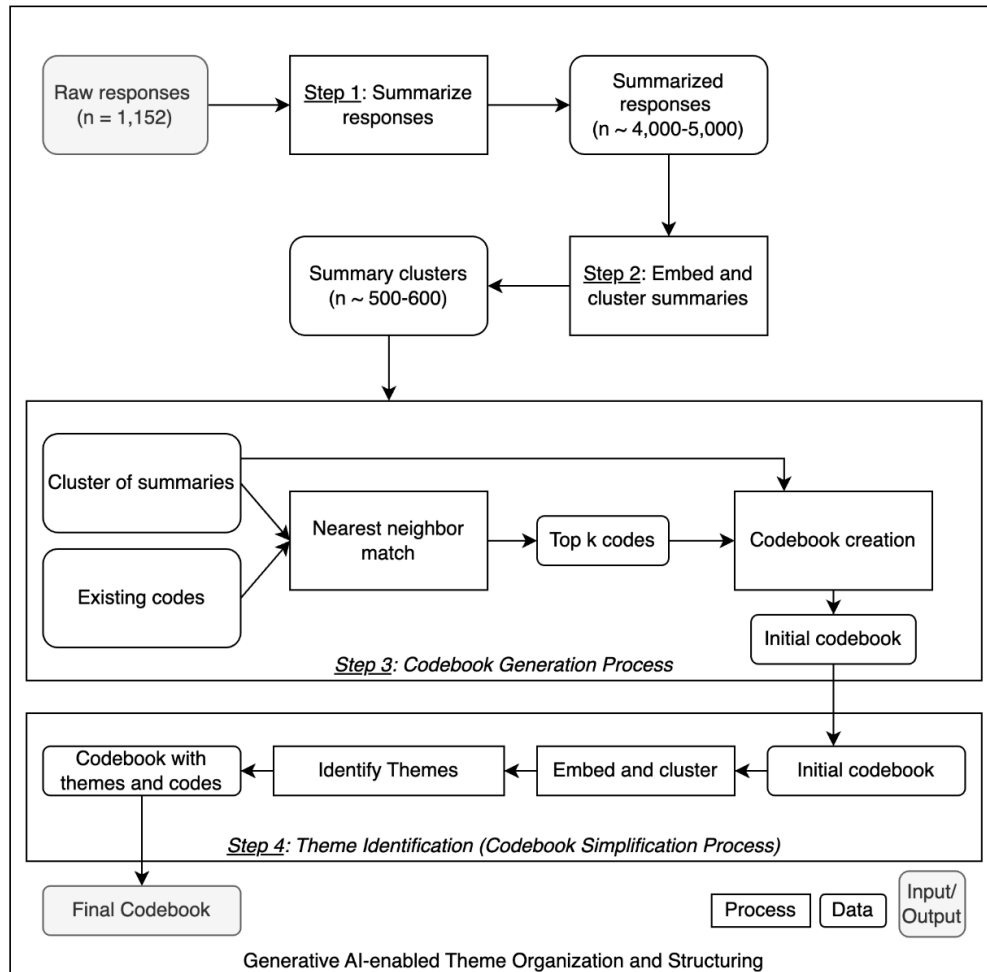


Figure 5: Workflow for the method

Figure 2: Workflow: GATOS pipeline drawing

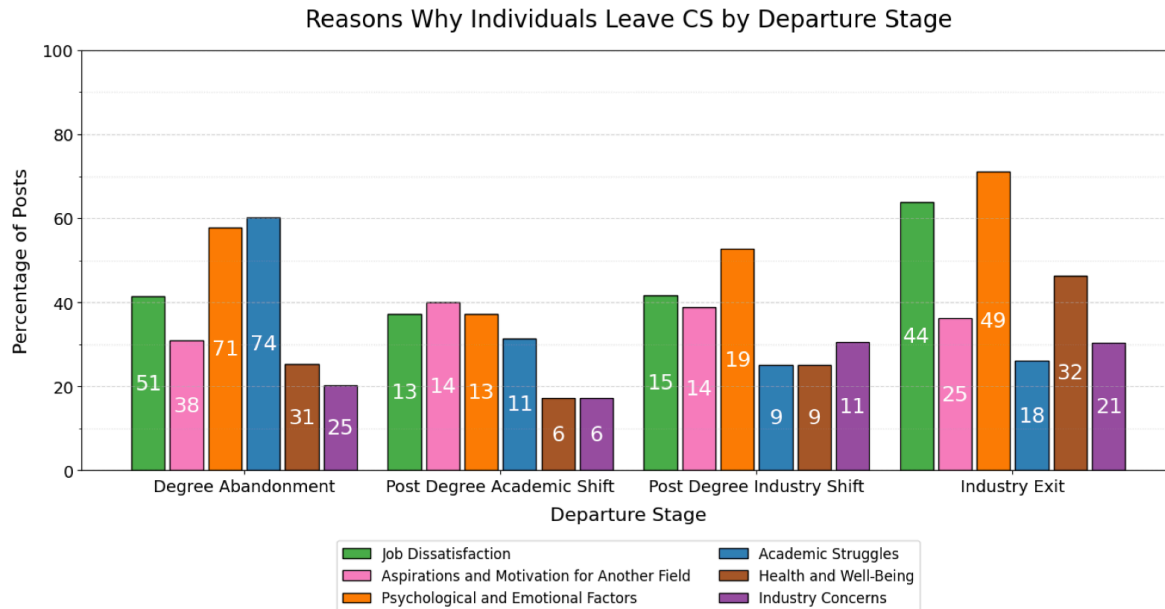


Figure 3: Why leave CS? (Ross & Katz, 2025)

Received: 1 November 2024 | Revised: 22 May 2025 | Accepted: 4 July 2025
DOI: 10.1002/jee.70024

RESEARCH ARTICLE

JEE
Journal of Experiential Education

Expanding possibilities for generative AI in qualitative analysis: Fostering student feedback literacy through the application of a feedback quality rubric

Katherine Drinkwater Gregg | Olivia Ryan | Andrew Katz | Mark Huerta | Susan Sajadi

Figure 4: Teammate comment quality (Drinkwater et al., 2025)

TABLE 3 Feedback quality rubric.

Feedback criterion	0	1	2
Contributions to group tasks (role fulfillment, task management, task execution, project contributions, work output) <i>Exclude general references to teammate behaviors if they are not connected to a specific project task</i>	No task described	General or vague description of how the team member contributes to group tasks, with minimal details	Specific or detailed description of tasks that the team member contributes to that imply their value to the team
Behavior (interpersonal dynamics, team engagement, communications, personal attributes) <i>Exclude mentions of contributions to specific project assignments or tasks</i>	No behaviors described	General or vague description of team member behaviors, with minimal details	Specific or detailed description of team member behaviors that imply their value to the team
Gap (recognition of a negative difference between the ratee's performance and an expected standard)	No gap identified	A gap is alluded to or briefly mentioned, but lacks specific details on how it compares to an expected standard	A gap is discussed with specific details that highlight or easily imply how the gap compares to an expected standard
Action (suggested steps to remedy gaps or improve performance)	No actions identified	General or vague description of an action with minimal details	Specific or detailed description of how a team member should address a gap in their performance

Note: The italicized text indicates guidance added to replicate the researchers' application rules for the LLM ratings.

Figure 5: Example rubric (Drinkwater et al., 2025)

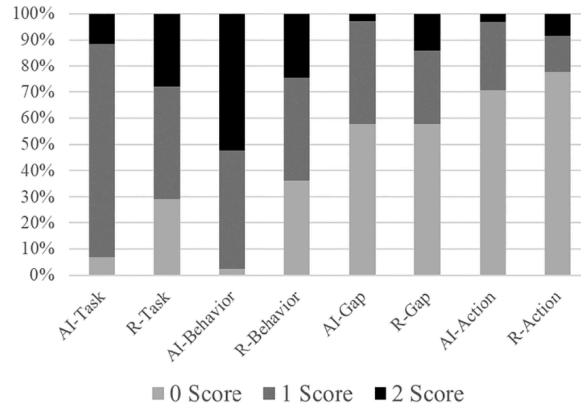


FIGURE 1 Comparison of Rubric Score Distributions for Feedback Quality by Researchers and LLM. The “AI-” prefix denotes the distribution from the LLM, while “R-” denotes researchers.

Figure 6: Results Table (Drinkwater et al., 2025)

Stumbling Our Way Through Finding a Better Prompt: Using GPT-4 to Analyze Engineering Faculty's Mental Models of Assessment

Figure 7: Using AI to Create Causal Influence Diagrams

Use Case 3

Part 1: Figurative Language Detection

10

How people construct mental models¹

Allan Collins & Dedre Gentner

Use Case 3

Part 1: Figurative Language Detection

- On learning and AI:
 - Pushing through mental barriers when you want to quit and when it's hard
 - The knowledge becomes embedded in your thinking
 - The illusion of learning from superficial digital content
- On AI policies:

- A legal and procedural minefield
 - Dual-level approach—similar to biosafety protocols
 - Competing with students is an arms race
-

Use Case 3

Part 1: 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
-

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

Rank	Model	Accuracy	Precision	Recall	F1	Backend
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
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

Figurative Language Results

Prompt format

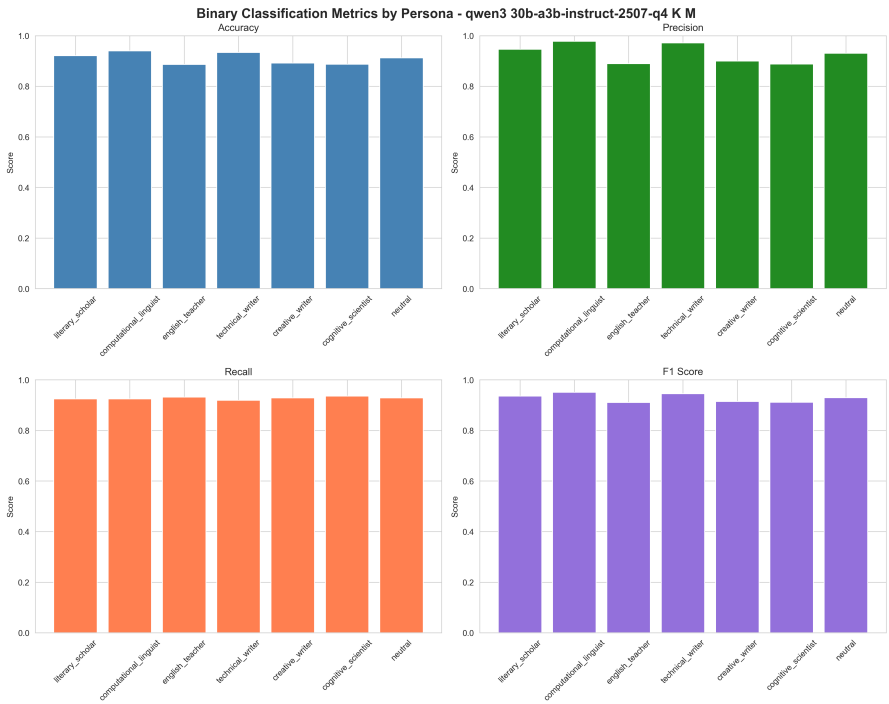
Format	Accuracy	Rank
All-Caps	87.7%	1 (tied)
Plain Text	87.7%	1 (tied)
Numbered	87.6%	3
Emoji	87.3%	4
Markdown	87.1%	5
XML	87.0%	6

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

Specific LLM Personas



Use Case 3

Part 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.”

Use Case 3

Part 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.”

Use Case 3

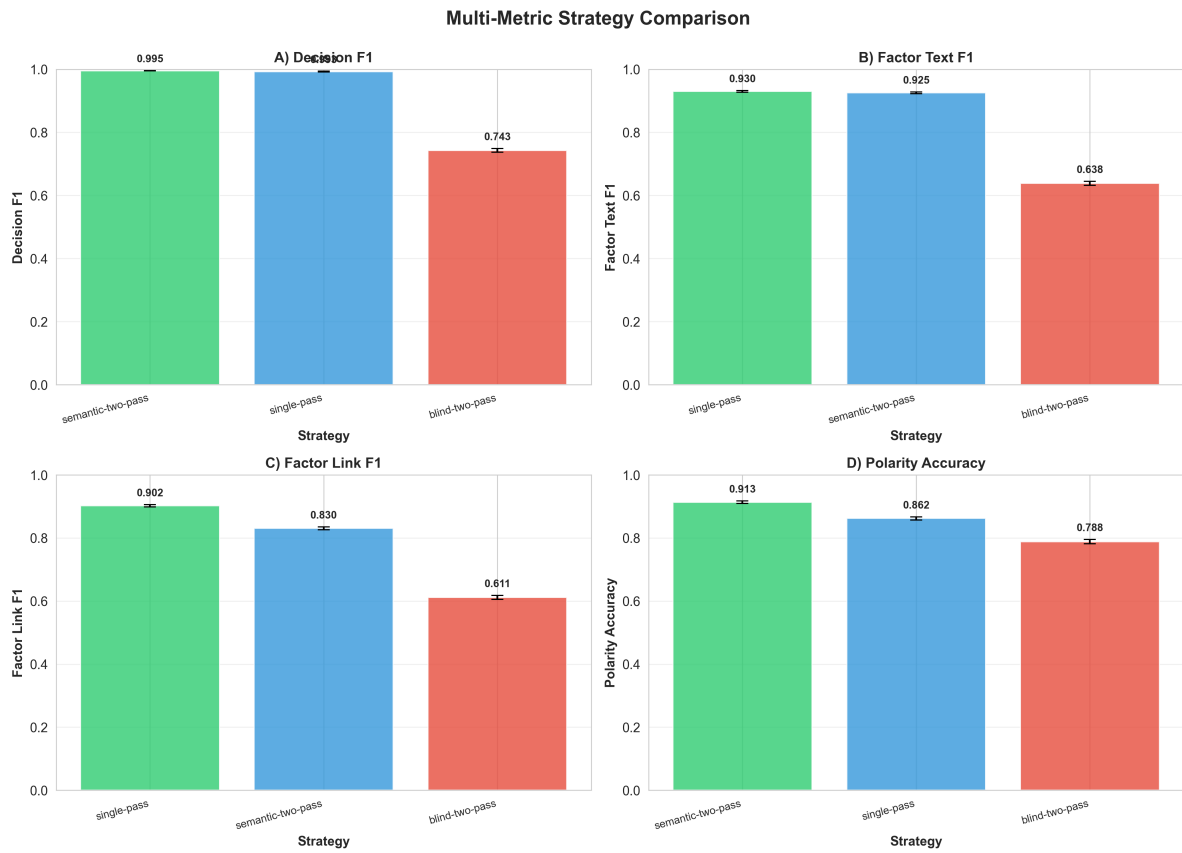
Part 2: Identifying Decisions and Factors

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

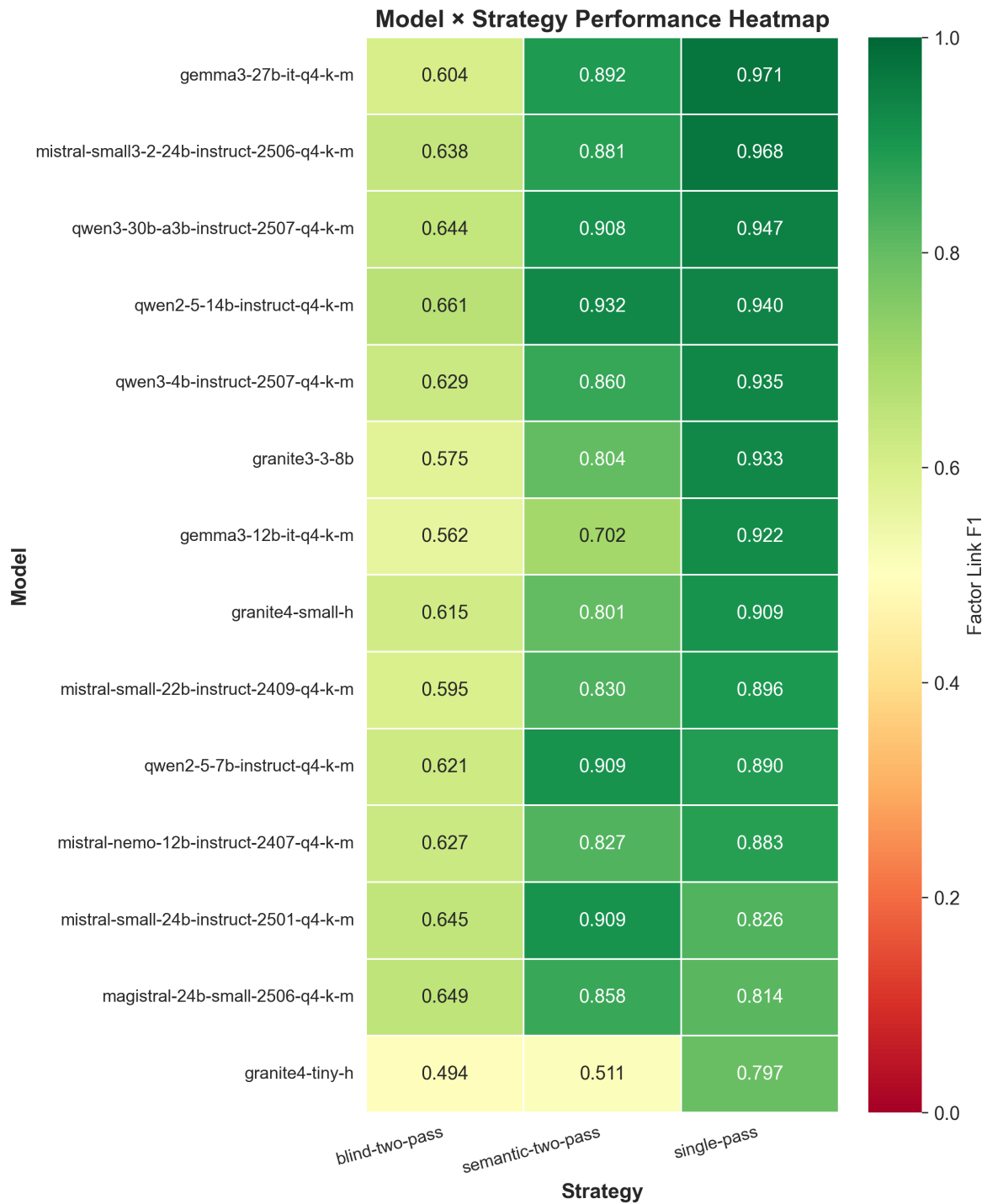
Architectures

- 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



Use Case 3

Part 3: Causal Beliefs

Brief Detour

Directed Acyclic Graphs (DAGs)

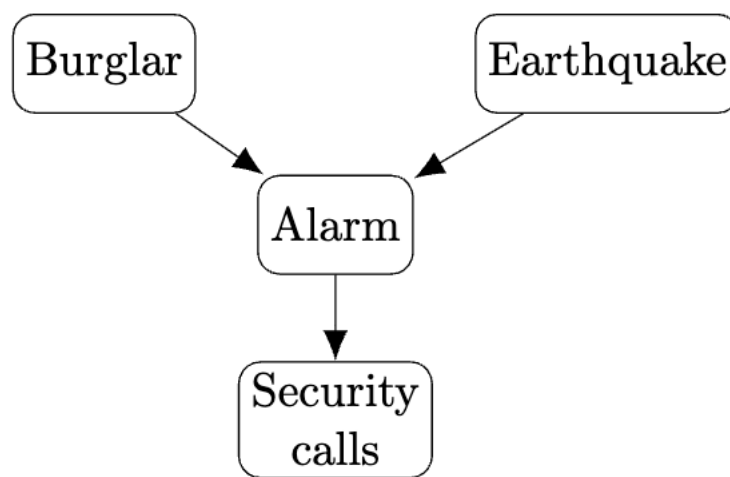


Figure 1.: An example of a causal graph ([Pearl, 2009](#)).

Figure 8: example from Pearl (2009)

Brief Detour

Causal Influence Diagrams

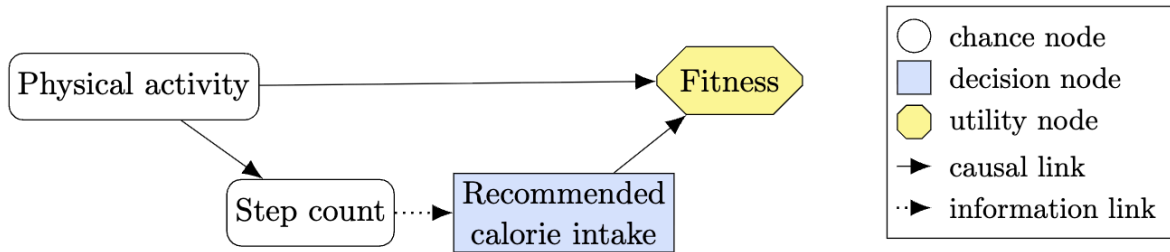


Figure 9: Simple example of causal influence diagram (Everitt et al., 2019)

Brief Detour

Causal Influence Diagrams

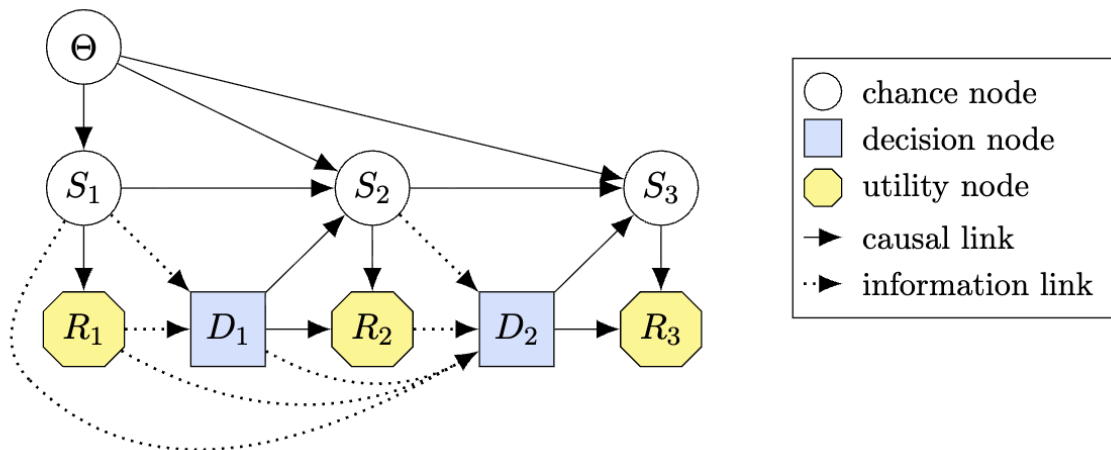


Figure 10: Complex example causal influence diagram (Everitt et al., 2019)

Use Case 3

Part 3: Causal Beliefs

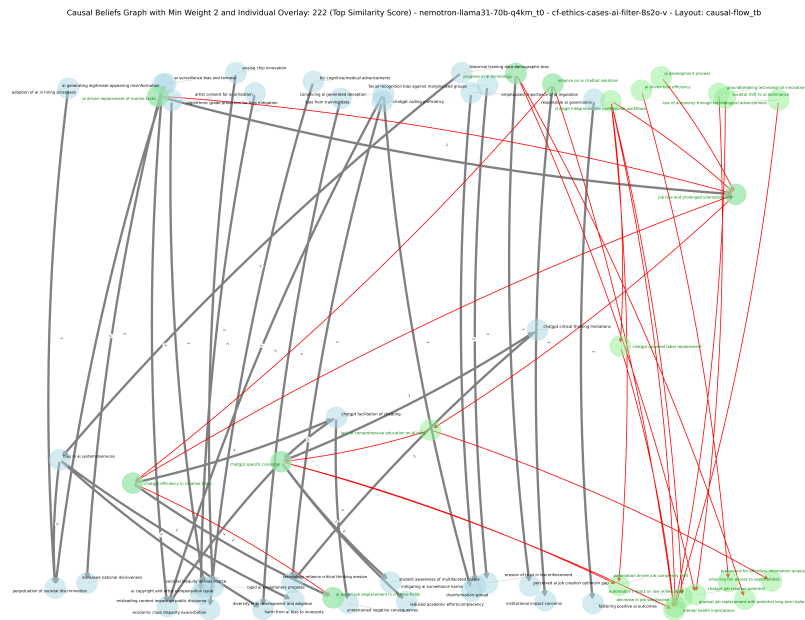


Figure 11: Example Inductive Causal Belief Network

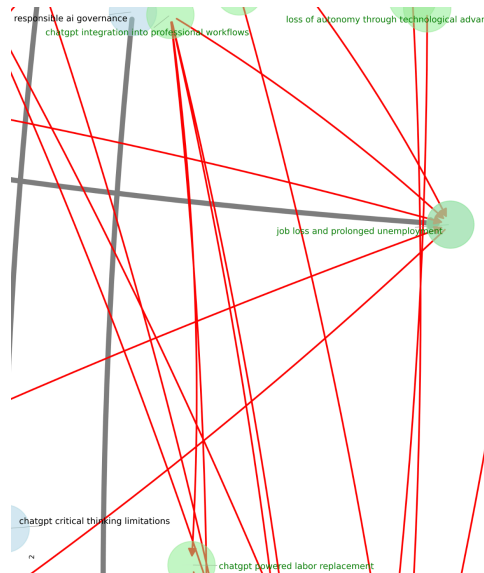


Figure 12: Closeup of Example Inductive Causal Belief Network

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

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] modified 43.55 t summarization 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>		
--	--	--

Figure 13: 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

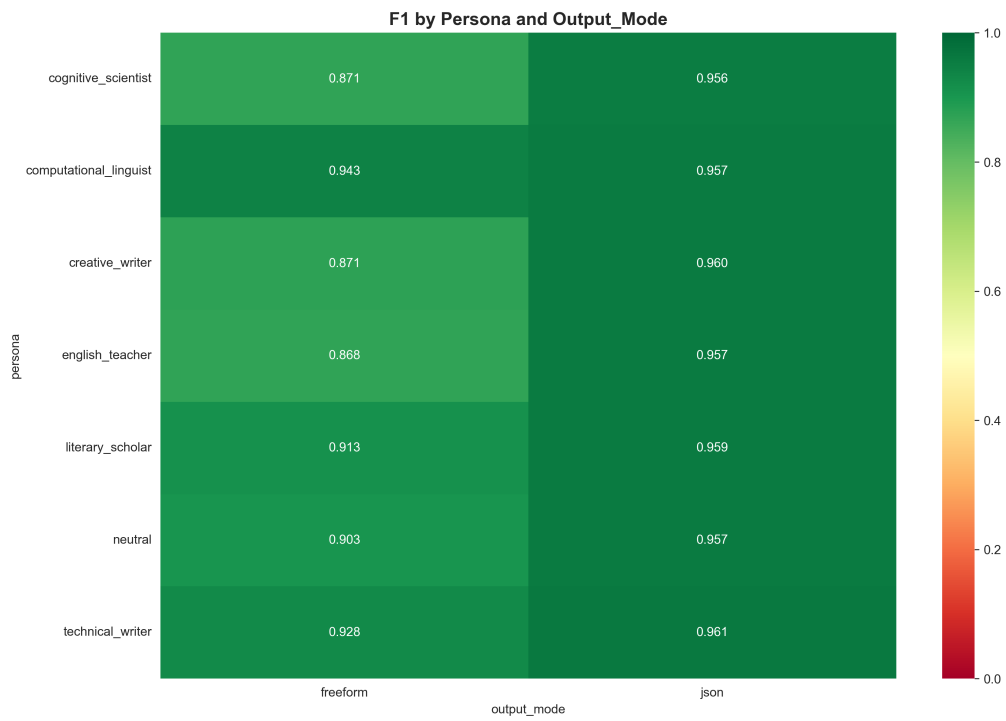


Figure 14: figurative language experiment results

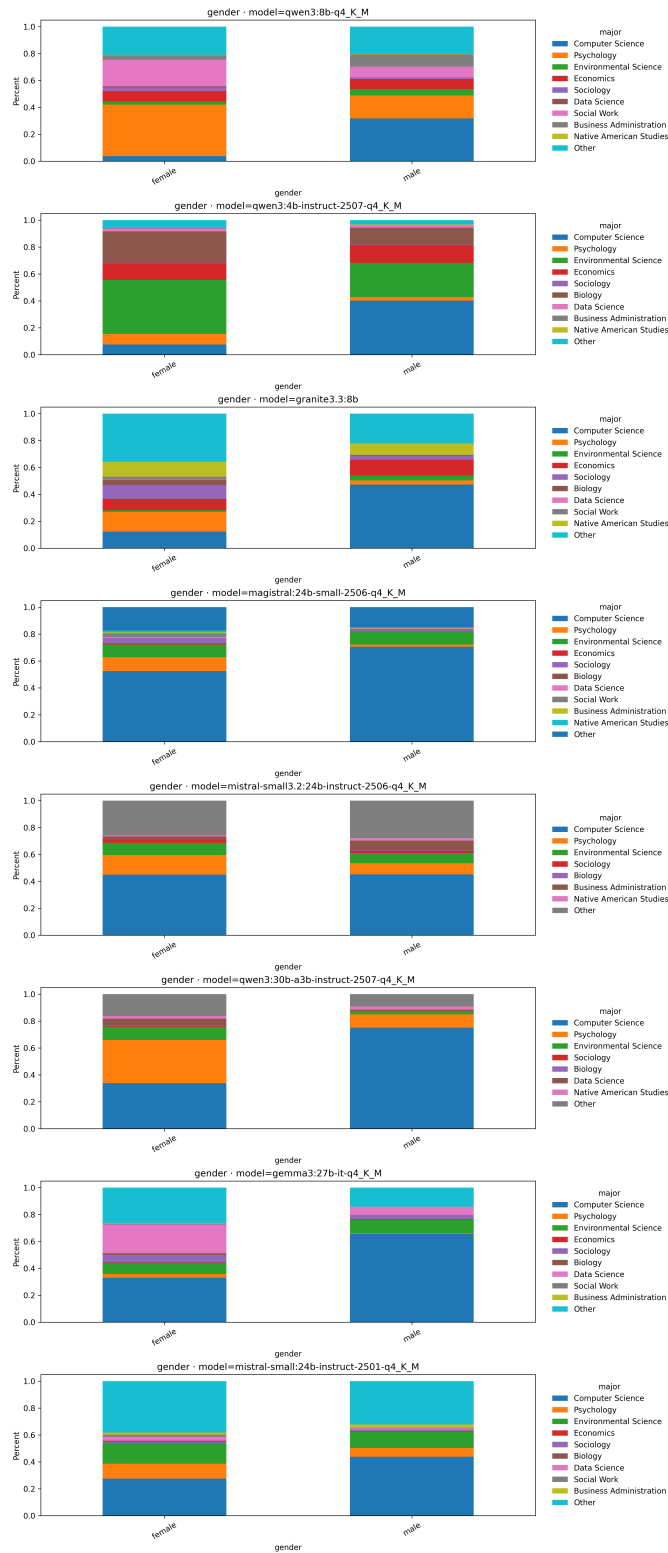


Figure 15: Major suggestions

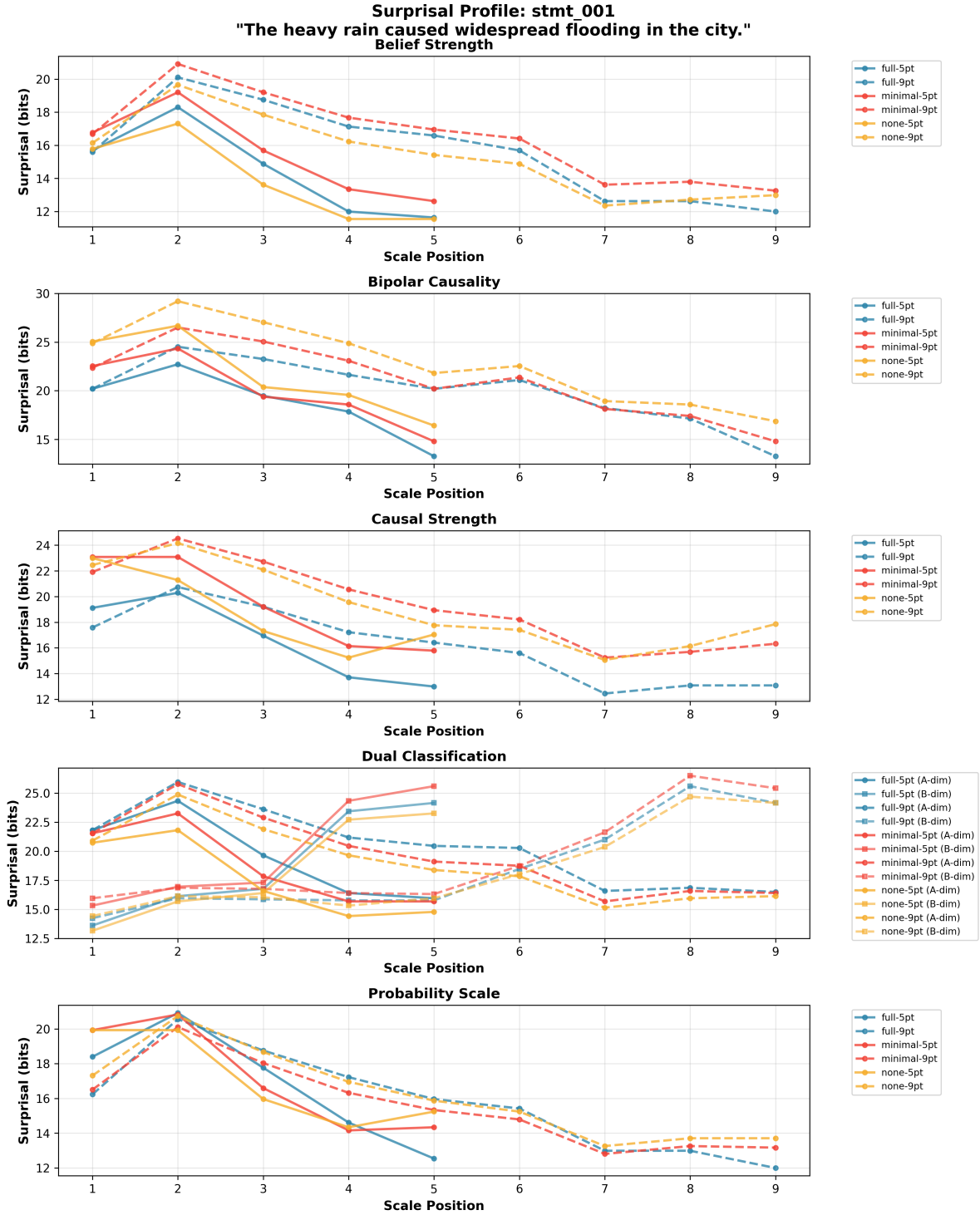


Figure 16: surprisal results for causal statements

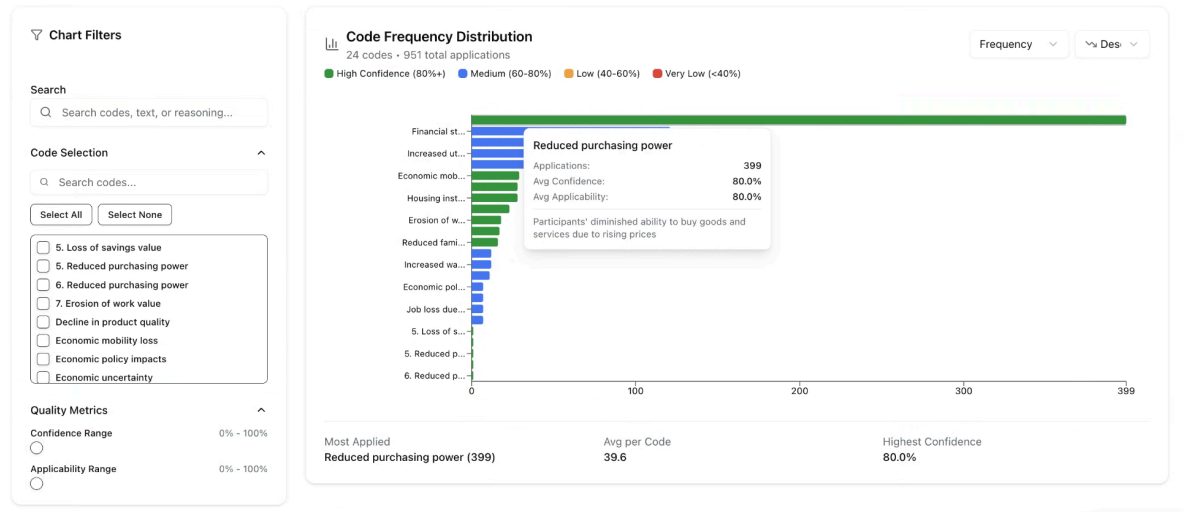


Figure 17: QDA platform

Online QDA platform

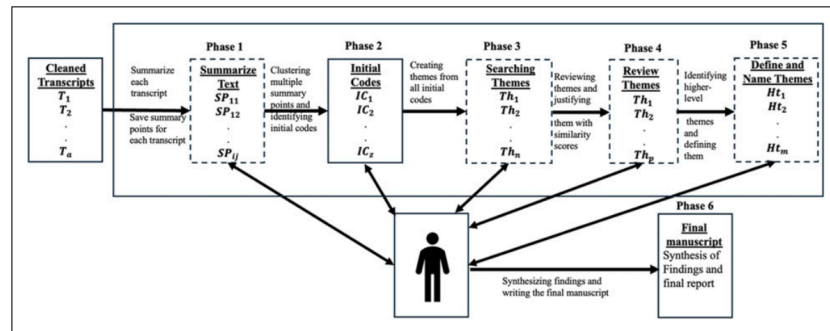


Figure 1. The Overall Workflow for Thematic Analysis With the NLP and GAI Technologies.

Figure 18: Workflow: human-in-the-loop