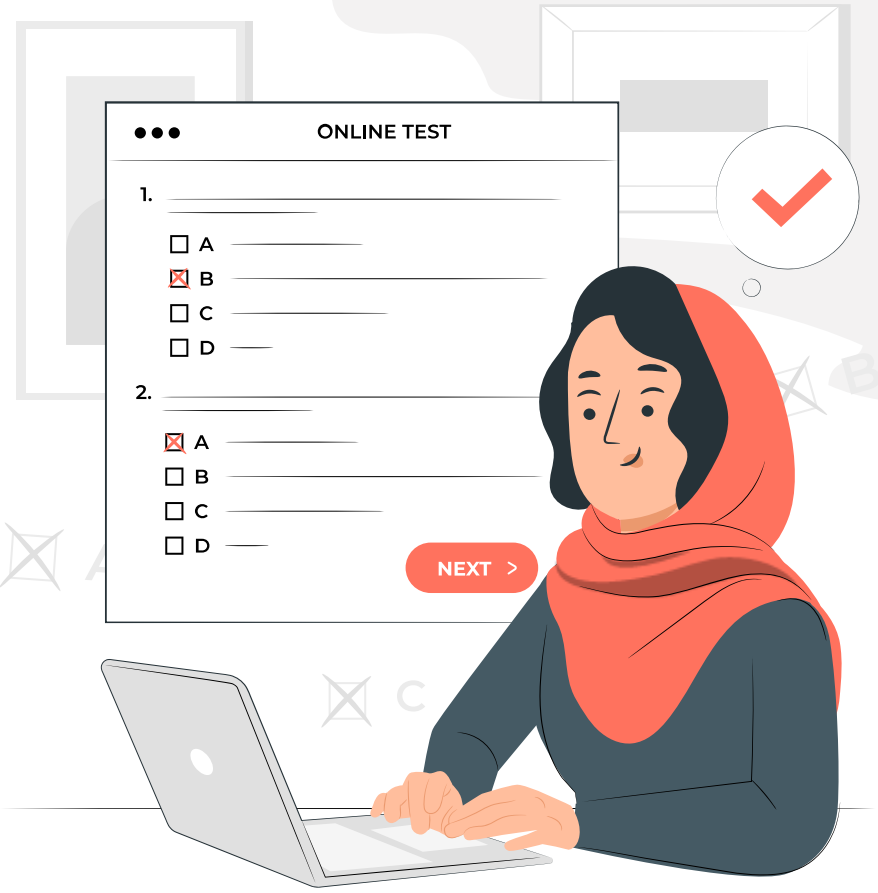
Generating Automatic Feedback for Open-Ended Questions with Fine-Tuned LLMs: A Comparison of GPT and Llama

Okan Bulut & Elisabetta Mazzullo

Measurement, Evaluation, and Data Science

University of Alberta

AIME – March 19, 2025



Selected-Response Items

Select an answer from a list

Multiple choice

True/False

Matching

Constructed-Response Items

Create your own answer

- Short-answer items (e.g., fill-in-the-blank)
- Extended-response items (e.g., essays)

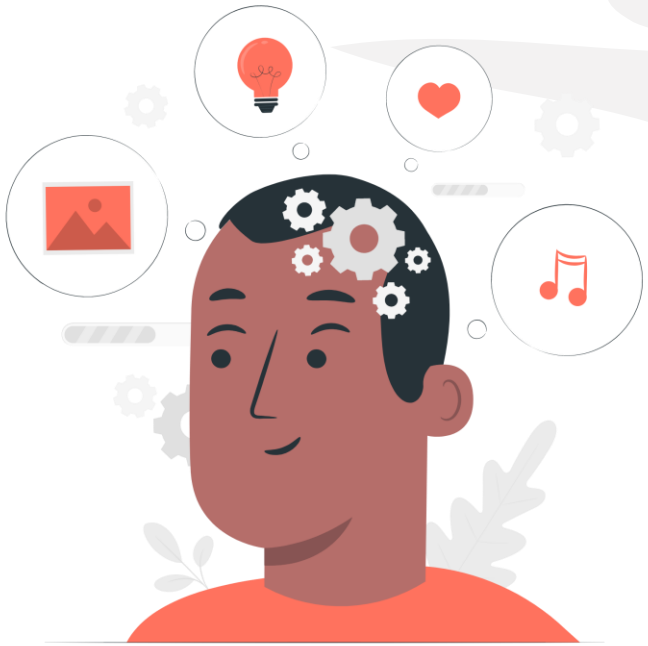
Feedback

Importance & Challenges

Feedback is most effective when:

- Timely
- Personalized
- Actionable[1]
- Balanced emotional tone[2]

Situational Judgement Tests



Unlike traditional achievement tests, SJTs measure various **non-cognitive** skills based on test-takers' actions for hypothetical real-life scenarios.

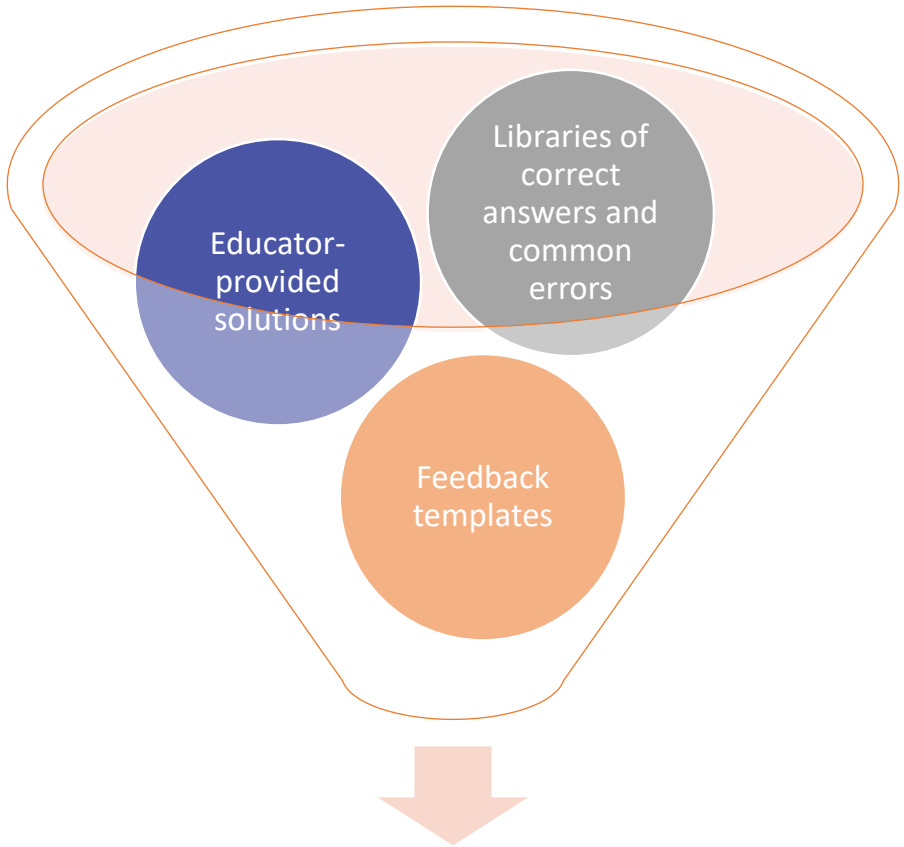
Casper as an SJT

A computerized test designed for assessing different aspects of **professionalism** such as ethics, empathy, and resilience.

CASPER

How Does Casper Work?

Automatic Feedback Generation (AFG)



Written/textual feedback, graphs, and dashboards (less personalized[3])

Data-driven AFG via pretrained large-language models (LLMs)

What We Know About LLM-Powered AFG

ChatGPT and Prompting for AFG

- Feedback generated via ChatGPT is close to that written by expert teachers
- Increases revision performance, task motivation, and positive emotions[4,5,6]
- High-quality prompting → high-quality feedback

Some issues

- Even with high-quality prompts, LLMs do not consistently satisfy user needs.
- Excessive reliance on prompting strategies can create a barrier to access.
- Lack of data privacy when proprietary models like ChatGPT/GPT are used

New Opportunities for LLM-Powered AFG

- **Parameter Efficient Fine-Tuning (PEFT)**
  - Adapts a pre-trained LLM to a downstream task with low memory and data requirements
  - Helps a model generate feedback better aligned with user preferences
  - Reduces heavy reliance on prompting
- **Beyond ChatGPT/GPT** 
  - Open-source LLMs, such as Llama from Meta AI, can provide:
    - Comparable performance to state-of-the-art GPT models
    - Higher privacy
    - Lower long-term costs

Our Study 

- **Objective:** To investigate the potential of fine-tuning pretrained LLMs to generate feedback for open-ended items

- We aim to answer the following questions:
  1. Can LLMs generate high-quality feedback for short open-ended responses?
  2. Can the open-source Llama model deliver feedback of comparable quality to proprietary GPT models?
  3. Does fine-tuning improve the quality of feedback generated by these LLMs?

Data

- **Main corpus:**
  - 211,058 written responses to 103 text-based scenarios in Casper
  - The list of soft skills assessed by each scenario (no universally acceptable right or wrong)
  - Scenario texts
  - Two sets of scoring guidelines
  - Scores assigned by human raters

- The first guideline, "**guiding background**," provides detailed context about the focal skills assessed, their relevance to the scenario, and how they should emerge in responses.

- The second guideline, "**guiding questions**," distills this background into three to four concise questions, such as "Did the applicant demonstrate [skill]?" or "Did they consider [topic]?"

Example 1: Self-Awareness

Forgiveness and compassion are always linked.

Q1. Briefly describe an experience where you demonstrated forgiveness or compassion.

Q2. What did you learn about yourself from this experience? Explain your response.

Q3. Do you typically find it more challenging to demonstrate forgiveness or compassion? Explain your reasoning.

Example 2
Motivation & Professionalism

You have been working at a job for six months. It is exactly your field of expertise, and you hope to begin a career in this field. However, you regularly find yourself without any work to do.

Q1. Would you talk to your supervisor about the lack of work in this situation? Why or why not?

Q2. Imagine that a co-worker was having a similar experience and felt demoralized. What advice would you give your co-worker to help to increase morale? Explain your reasoning.

Q3. In workplace settings, are you typically more focused on working toward your own personal goals or the organization's goals? Explain your response.

Sample Responses

Scenario 1, Low Score

When a friend **canceled plans last minute**, but I forgave them and didn't hold it against them. I learned that I can be understanding when I know someone is strugling, and I don't let small things affect our relationship.

I find it **more challenging to be compassionate** sometimes, especially if the person is not showing any remorse. it's harder to be compassionate if I don't understand where they're coming from.

Scenario 1, Low to Moderate Score

One time, a friend was really **rude to me when I asked for help** with homework. It hurt me, but later I found out she was going through personal issues. Even though she didn't apologize, I chose to forgave her.

I learned that it's important to try and understand where others are coming from, especially when they've hurt me. It's not always easy, but *I realized that forgiving helps me feel better and move on from the situation.*

I think **both forgiveness and compassion can be difficult**, depending on the situation. If someone hurt me badly and didn't apologize, I would find it harder to be compassionate. But I would still try to forgive, even if it's tough, because I know it's better for my peace of mind.

Sample Responses

Scenario 1, High Score

A situation where I demonstrated compassion was when a close **friend was going through a difficult time** due to a family isue. While *initially, I didn't fully understand* her actions and reactions during that period, I made an *effort to listen without judgment* to her. I knew she was struggling, and rather than getting frustrated with her behavior, I tried to show understand. I o*ffered her my support by* being there for her, even when she seemed distant. *By doing so, I hoped* to show her that I cared, even if I couldn't immediately fix the situation.

This experience taught me that I have the capacity for empathy, even in challenging situations. I realized that I am able to be patient with others and *offer support, even when I might not fully understand what they are going through*. I learned that *showing compassion doesn't always require fixing the problem but simply being presen*t for someone in their time of need. It also helped me understand that it's not always about me or my comfort level; it's about meeting the other person where they are and offering kindness without expecting anything in return. **forgiving** someone can be difficult for me because it requires letting go of feelings of hurt or anger.

Creating Training Data

- No predefined examples of “ideal feedback” were available.
- OpenAI suggests that “50-100 examples are often sufficient to see clear improvements in GPT models”.
- The AFG literature is also congruent with this claim:
  - A small, high-quality training dataset can yield strong performance in feedback generation[7].
- Feedback samples were crafted for 124 responses to 12 Casper scenarios.
  - Broad representation of scores and skills across the scenarios

How We Defined “Ideal Feedback”

Feedback statements…	were written in the **second person** to foster a personal and engaging tone[8].
	balanced **positive and negative** aspects[9].
	maintained a **supportive** tone.
	offered **actionable** suggestions.
	integrated **scoring guidelines.**
	addressed the **unique qualities** of each response, further enhancing its relevance and personalization[10].

Pre-trained LLMs

GPT-4o-mini

- Decoder-only transformer
- Parameters: 8 billions (estimated)
- Context window: 128K tokens

Llama-3.2-11B

- Decoder-only transformer
- Parameters: 11 billions
- Context window: 128K tokens

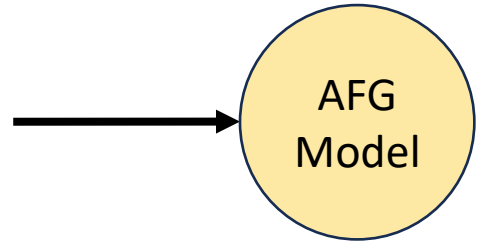
Model Training

Fine-tuning the pre-trained LLM on a collection of open-ended questions using instructions

Prompt engineering

One-shot prompting: giving the model one example of desired output

Chain-of-Thought prompting: asking the model to think step-by-step



Prompting LLMs

Four elements for effective prompting: context, question, format, and examples.

Context

The Casper test is a high-stakes situational judgment test that measures ten social intelligence skills. You are a tutor who helps applicants prepare for this test. Your job is to provide feedback that is grammatically correct, pertinent, personalized, actionable, and supportive.

Format

Write your feedback as if you were speaking directly to the applicant. Use the provided item, evaluation criteria, response, and score to guide your feedback. Lead applicants to improve their response without revealing the criteria too explicitly. Keep the feedback concise, ideally no more than 4 -5 sentences, and maintain an encouraging tone.

One-shot prompting

This is one example of the task and how you should solve it. Context: [item]. Base your feedback on the following criteria: [evaluation guidelines]. Disregard spelling, grammar and style. When you generate feedback , let's think step by step. Explain your reasoning process and how your feedback relates to the answer and the evaluation criteria. Response: [response]. This answer got a score of [score] out of 9. Feedback: [human- written feedback].

Chain-of-Thought prompting

Now, generate feedback for the answer to the following questions: [item]. Base your feedback on the following criteria: [evaluation guidelines]. Disregard spelling, grammar and style. When you generate feedback, let's think step by step. Explain your reasoning process and how your feedback relates to the answer and the evaluation criteria. Response: [response]. This answer got a score of [score] out of 9.

Example Feedback

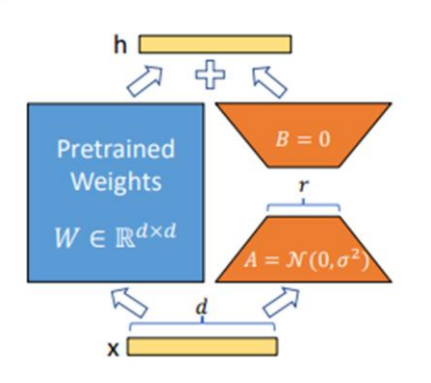
Your response expresses a partial agreement with the statement about forgiveness and compassion, highlighting the importance of understanding and empathy in forgiveness. However, the explanation lacks depth and could benefit from providing concrete examples or personal experiences to support your perspective. The factors you mention, such as envy and anger, are valid challenges to compassion, but your response lacks strategies or insights into overcoming these challenges. Furthermore, your statement about the link between self-happiness and compassion is somewhat vague and could be strengthened by offering more specific insights or examples. In discussing the relevance of compassion in healthcare, you touch upon its importance but don't delve into specific ways it can impact patient care or the healthcare environment. To improve your response, consider providing more detailed examples and expanding on the importance of compassion in healthcare, demonstrating a deeper understanding of its role in your future career.

Personalization Positives Flaws Actionable suggestions

Fine-Tuning LLMs

Llama

Quantum Low Rank Adaptation (QLoRa)[8]



Instruction

Context + Format

Question with Chain-of-Thought prompting

GPT

Fine-tuned through the OpenAI API

Also voiced to use QLoRA for fine-tuning

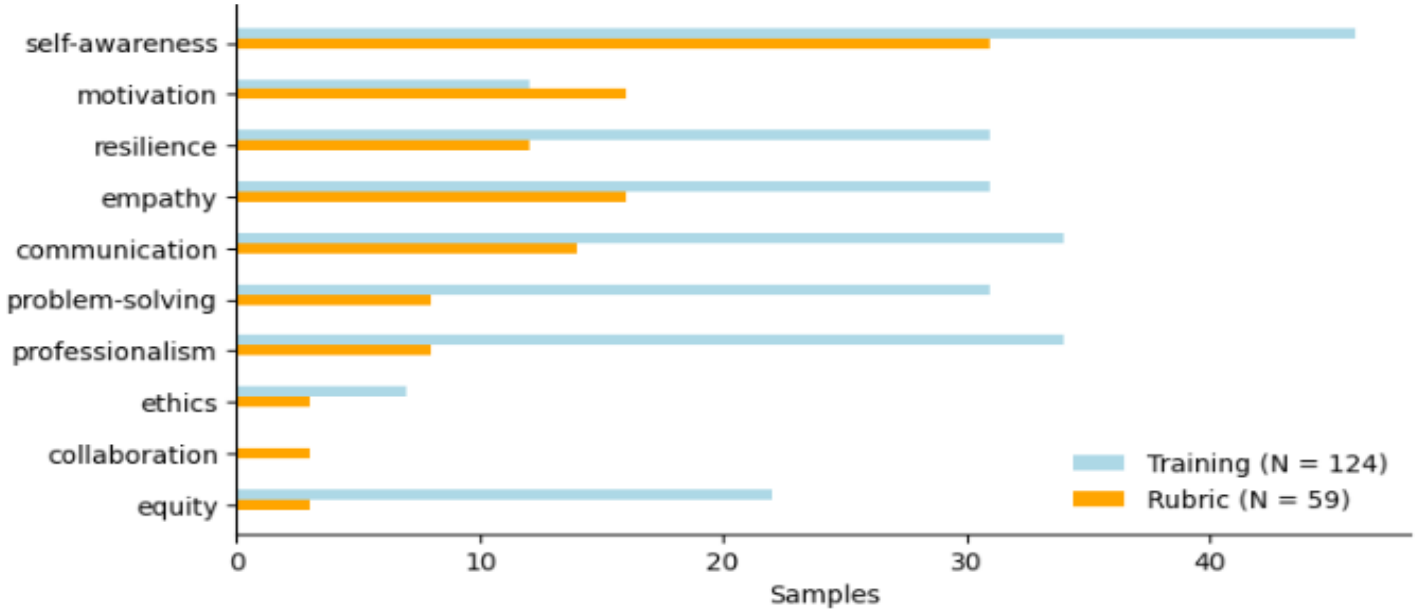
Instruction

Context + Format

Question with Chain-of -Thought prompting

Model Evaluation

- 59 feedback messages generated by the AFG model on unseen Casper responses
  - 6 responses for each score between 1-5
  - 5 responses for each score between 6-9



Evaluation via Rubric

- Outputs from each model were evaluated by two independent judges based on six structure-related criteria.

- Perfect or close-to-perfect agreement

Linguistic Quality	Correct Spelling Syntax Semantic
Factuality	Pertinent Response Scenario
Personalization	Personalized Generic
Actionability	Actionable Not actionable 9 N/A 9 actionable
Affective tone	Balanced Suggestions Flaws 9 not applicable 9 suggestions
Second Person	Yes No

Evaluation via Automatic Metrics

Polarity

-1 = extremely negative feedback

0 = neutral feedback

+1 = extremely positive feedback

Lexical Similarity

The extent to which two passages share characters, words, or n-grams

Reference text: Fine-tuned GPT-4o-mini

BLEU, ROUGE based on 1-grams, and the METEOR score

Values closer to 1 indicate higher similarity

Semantic Similarity

SentenceBERT to obtain feedback embeddings

Cosine similarity to compute semantic similarity between embeddings

- -1 = opposite meanings
- 0 = no similarity
- +1 = identical meaning

Results (Rubric-Based)

		GPT-4o-mini	Llama-3.2-11B	Ft:GPT-4o-mini	Ft:Llama-3.2-11B*
Linguistic Quality	Correct	100	100	79.7	98.3**
	Spelling	0	0	15.3	0
	Syntax	0	0	3.2	0
	Semantic	0	0	1.7	0
Factuality	Pertinent	100	96.6	94.9	98.3
	Response	0	1.7	5.1	0
	Scenario	0	1.7	0	0
Personalization	Personalized	96.6	100	98.3	86.4
	Generic	3.4	0	1.7	11.9
Actionability	Actionable	89.8	89.8	86.4	86.4
	Not actionable	0	0	3.4	1.7
	9 N/A	5.1	1.7	10.2	5.1
	9 actionable	5.1	8.5	0	5.1
Affective tone	Balanced	1.7	30.5	30.5	33.9
	Suggestions	88.1	66.1	54.2	49.2
	Flaws	0	0	5.1	5.1
	9 not applicable	5.1	1.7	10.2	5.1
	9 suggestions	5.1	1.7	0	5.1
Second Person	Yes	100	52.5 + 39***	98.3	96.6
	No	0	8.5	1.7	1.7

Table 2: Proportion of Feedback Meeting the Qualities of Effective Feedback Considered in the Rubric
Note. Ft: stands for fine-tuned [a] [b] [c]

[a]* percentages do not add up to one because one output (1.7%) was the repetition of part of the prompt and could not be evaluated to fit into any of the rubric categories.

[b]** one output (1.7%) was linguistically correct but repeated the same sentence twice.

[c]*** 39% of outputs provide second-person feedback, but only after analyzing the response in the third-person.

Results (Automatic Metrics)

Polarity

Model	Mean (SD)	Range (min, max)
GPT-4o-mini	0.21 (0.09)	0.02, 0.48
Llama-3.2-11B	0.01 (0.08)	0, 0.6
Ft: GPT-4o-mini	0.18 (0.10)	-0.06, 0.46
Ft: Llama-3.2-11B	0.14 (0.13)	-0.22, 0.4

Correlation between Scores and Polarity of Feedback

Model	Correlation
GPT-4o-mini	0.21
Llama-3.2-11B	-0.20
Ft: GPT-4o-mini	0.08
Ft: Llama-3.2-11B	0.37

Results (Automatic Metrics)

Average Lexical Similarity with Feedback Generated by Ft:GPT

	BLEU	ROUGE	METEOR
GPT-4o-mini	0.13	0.16	0.11
Llama-3.2-11B	0	0.03	0.01
Ft: Llama-3.2-11B	0.14	0.18	0.13

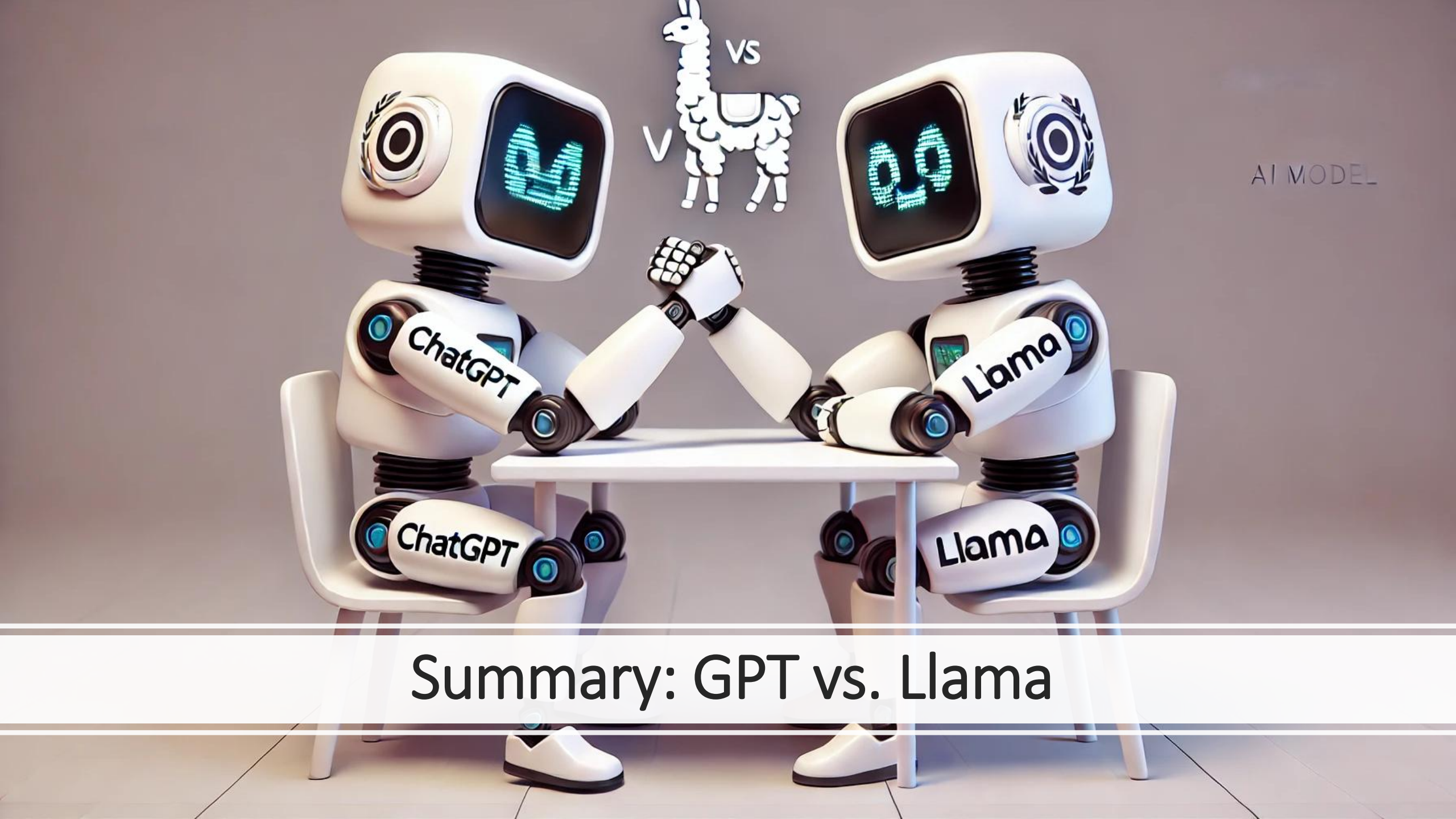
Average Semantic Similarity between Feedback Generated by the Four Models

	GPT-4o-mini	Llama-3.2-11B	Ft:GPT	Ft:Llama
GPT-4o-mini	1			
Llama-3.2-11B	0.30	1		
Ft:GPT	0.54	0.27	1	
Ft:Llama	0.55	0.29	0.53	1

Results (Automatic Metrics)

Average Semantic Similarity of Feedback with Evaluation Guidelines and Responses

Model	Evaluation Guidelines	Response
GPT-4o-mini	0.52	0.38
Llama-3.2-11B	0.20	0.16
Ft: GPT-4o-mini	0.50	0.38
Ft: Llama-3.2-11B	0.52	0.38



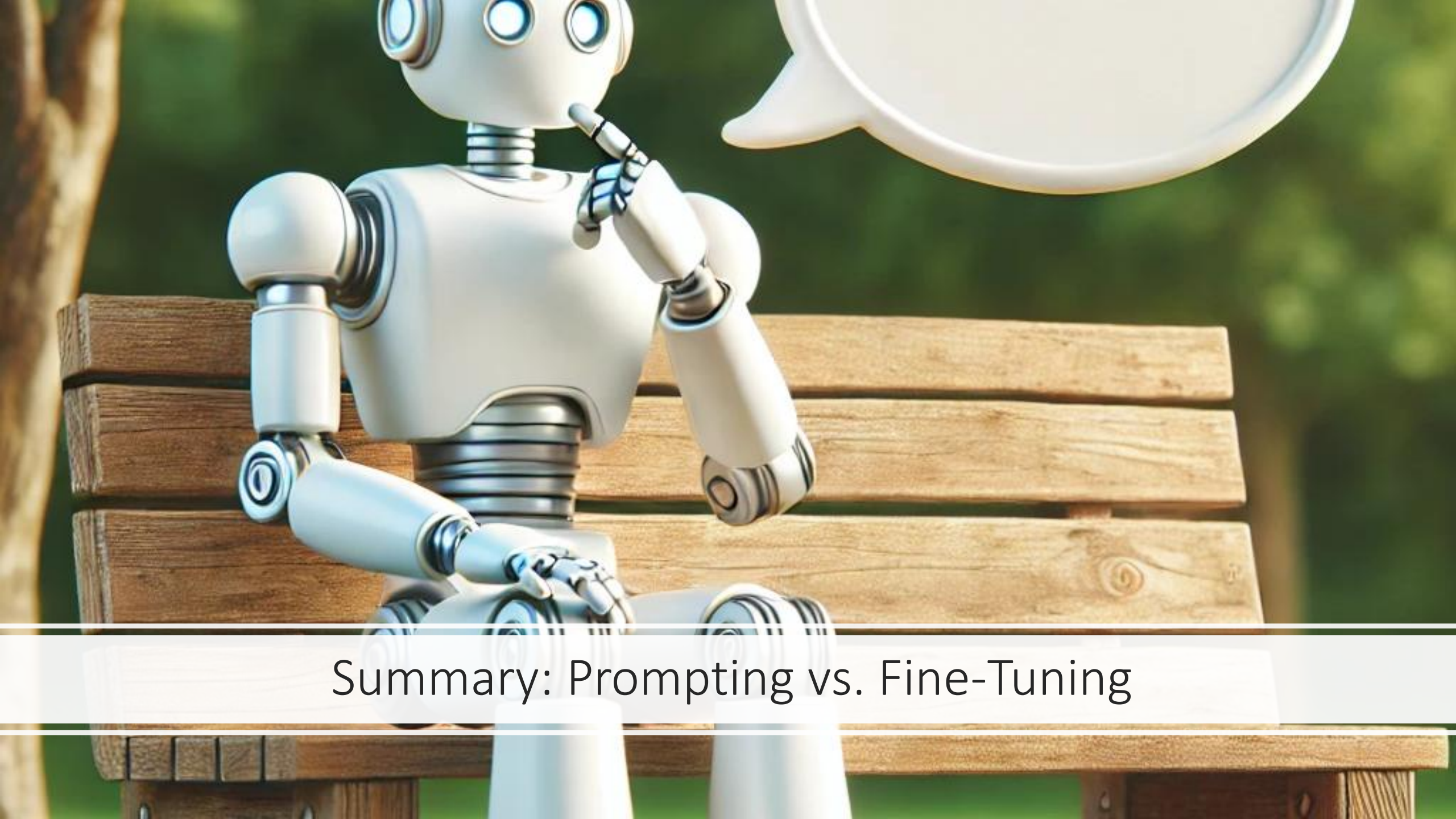
Summary: GPT vs. Llama

- Both GPT and Llama provide feedback that generally shows the quality of effective feedback.

- Although prompted with the same instruction, the two LLMs generated feedback with different structures and styles.

- GPT tends to give **more positive feedback**. However, feedback has a very positive tone, even for low-score responses.

- Compared with base Llama, GPT outputs tend to be more similar to those generated by our fine-tuned GPT, to applicants' responses, and to the evaluation criteria.



Summary: Prompting vs. Fine-Tuning

👍 After fine-tuning,

- GPT learns to give more balanced feedback.
- Llama
  - learns to adopt a style more similar to that seen in the training examples.
  - generates feedback with higher similarity to responses and evaluation guidelines.

👎 After fine-tuning,

- GPT's linguistic abilities deteriorate.
- Fine-tuned Llama does not provide personalized feedback as consistently as base Llama.
- Both LLMs provide actionable feedback less frequently.

- Fine-tuning did **not** show a clear advantage over carefully prompting foundational LLMs
  - Fine-tuning LLMs on larger, high-quality datasets may still offer benefits.

Key Takeaways

- How to create high-quality prompts for feedback generation is still uncharted territory.
- Using GPT or Llama, high-quality prompting is generally sufficient to generate outputs that largely meet the qualities of effective feedback.
- Fine-tuning seems mostly successful in helping the model adopt a style similar to that observed in the training examples.
- Educational stakeholders should get involved in the development of larger training datasets and the evaluation of automatic feedback.

Limitations and Future Research

- A small training dataset (*N* = 124 feedback samples)
  - The size of the training dataset may be modified to examine its impact on the generated feedback.
- We don’t know whether the generated feedback would improve test-taker performance
  - What type of feedback is the best for test-takers is still a mystery...
  - Broad feedback to guide test-takers’ future attempts vs. specific feedback based on the answered questions

THANK YOU!

For questions and comments:

Okan Bulut
bulut@ualberta.ca

Elisabetta Mazzullo
mazzullo@ualberta.ca