

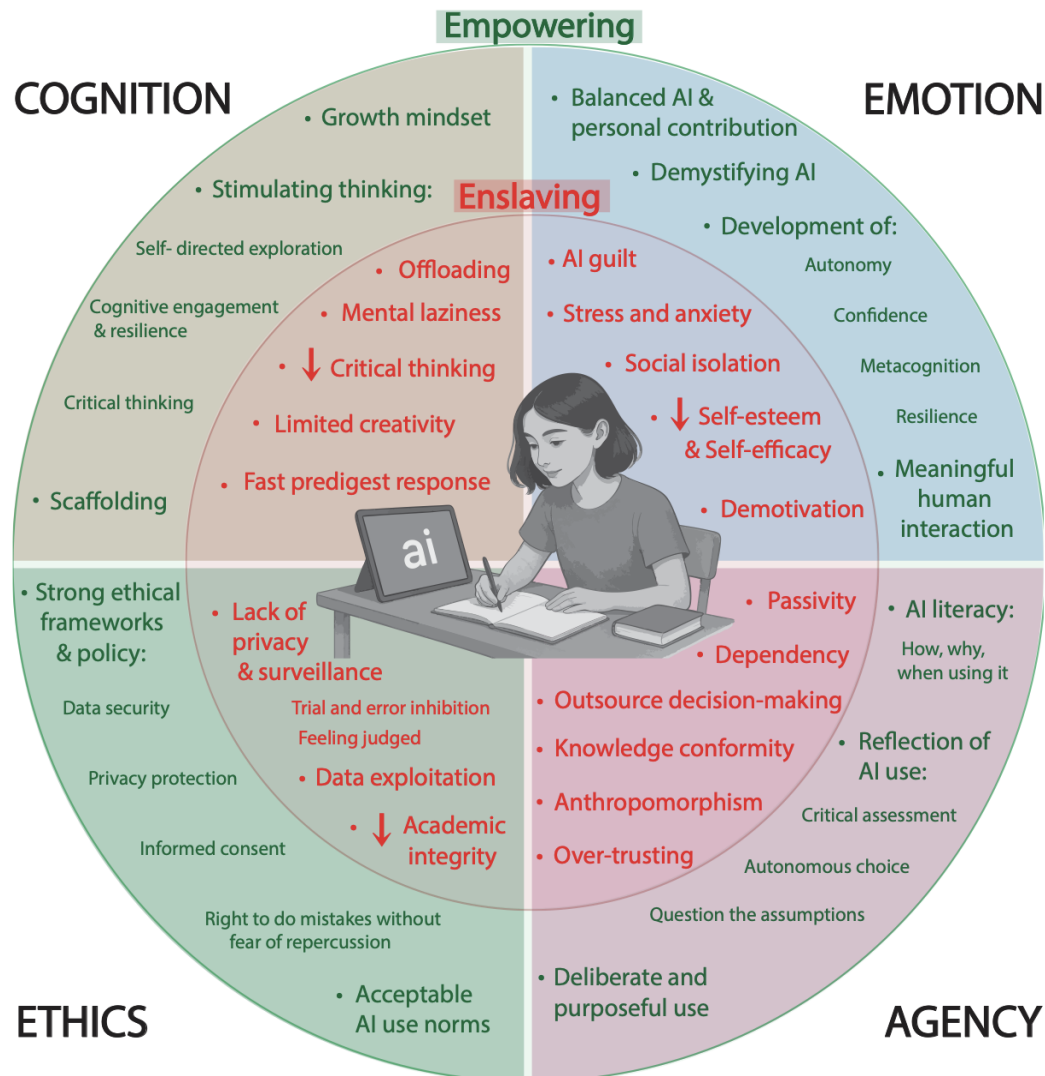


An LLM-based educational chatbot to promote Critical Thinking

Lucile Favero Montero

Co-advisors: Nuria Oliver, Juan Antonio Pérez-Ortiz & Tanja Käser

Toward a critical use of AI in education



Increasing integration of AI tools in education

- Empowering:**
- Support learning
- Challenges:**
- Cognitive atrophy
 - Loss of agency
 - Emotional risks
 - Ethical concerns



Rather than allowing chatbots to **weaken our cognitive abilities**, my objective is to develop a chatbot that **supports the enhancement of critical thinking and self-regulation skills**

A Socratic approach to an educational chatbot



Traditional chatbot

Provides the knowledge to facilitate the learning



Socratic chatbot

Asks specific and targeted questions to help the learners build knowledge by themselves

A Socratic approach to an educational chatbot

Aims

- Reflexion
- Purposeful learning
- Critical thinking

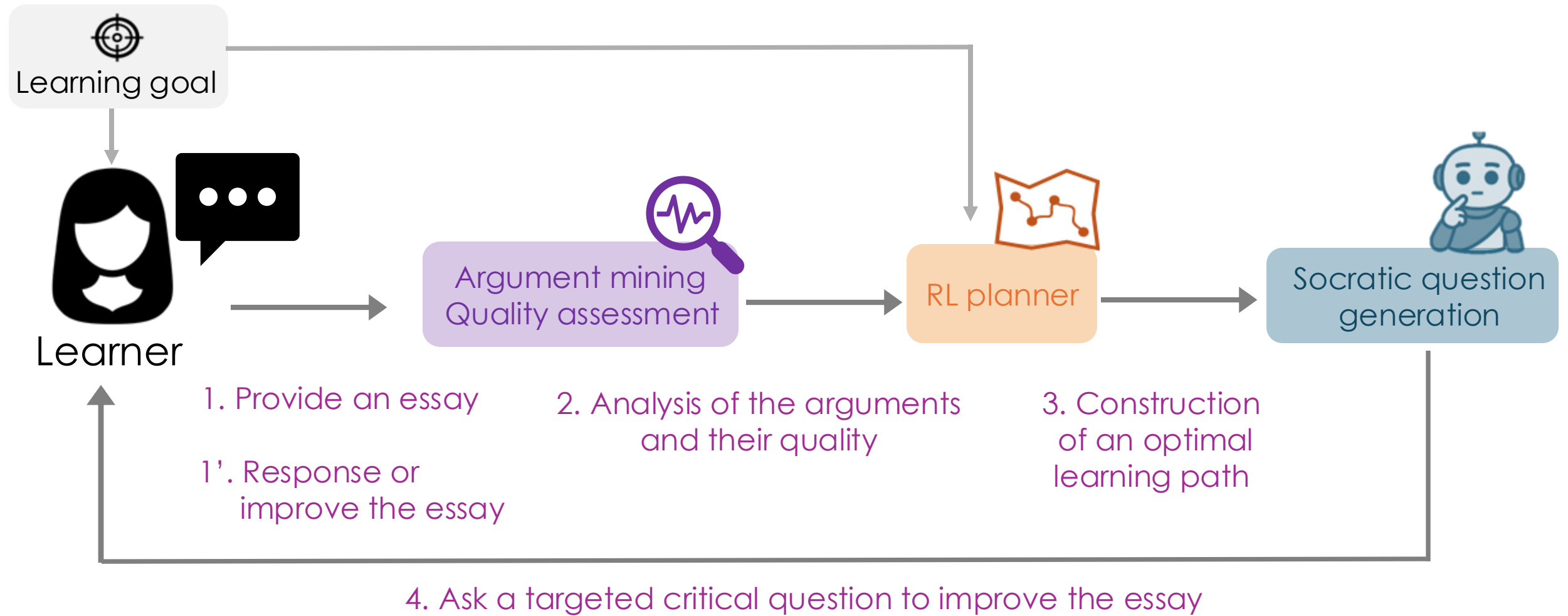
Educational Tasks

- Question answering
- Self-regulation
- Knowledge discovery

Broader applications

- Combatting fake news
- Promoting democracy

A Socratic AI educational chatbot Maïke



Socratic questioning

What is Socratic questioning?

Socratic questioning is a **disciplined questioning** method used to **encourage critical thinking**, uncover underlying beliefs, and explore complex ideas

Socratic questioning

Taxonomy of Socratic questions based on Critical Thinking concepts¹

Question type		Description	Exemplar
01	Clarification	Question probing the ambiguities of a thought	What do you mean by ... ?
02	Probing assumption	Question probing the assumptions behind a thought	Why do you assume ... ?
03	Probing reasons and evidences	Question probing the justifications or concrete evidences that could have supported a thought	How did you know that ... ?
04	Probing implications and consequences	Question probing the impacts or implications of a thought	If ... , what is likely to happen as a result?
05	Probing alternative viewpoints and perspectives	Question probing other possible viewpoints	What else should we consider about ... ?

1. R. Paul, L. Elder, *The Thinker's Guide to Socratic Questioning*, Rowman & Littlefield, 2019.



Enhancing critical thinking in education by means of a Socratic chatbot



OBJECTIVE

Building a Socratic chatbot:

Generation of Socratic questions by prompt and fine-tuning a pre-trained LLM



METHOD

1. **SocratiQ**: Dataset of annotated Socratic questions and answers using r/changemyview



2. Fine-tuning **Llama2 7B and 13B four-bit quantized** with unsloth



3. Prompt-tuning of the fine-tuned model on Ollama



EVALUATION

Compare the learner's response and the ground truth

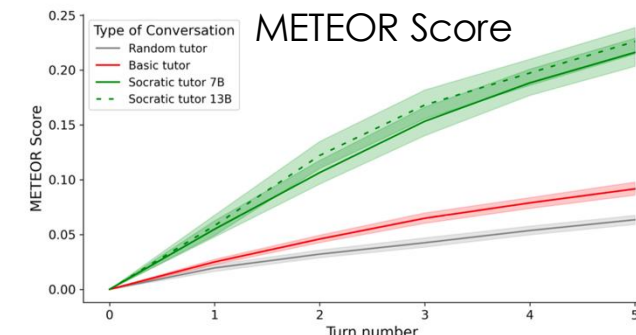
- *Generation of synthetic conversations*: prompt-tuning Llama 2 7B to simulate learner responses
- *Baselines*: 2 non-Socratic prompt-tuned tutors
- *Evaluation Dataset*: Theory of Knowledge International Baccalaureate questions



RESULTS

The Socratic tutors outperform the baselines on all metrics

Tutor type	BLEU	ROUGE-L	METEOR	BERTScore	LLM Score
Socratic Llama2 13B	3.65	0.157	0.226	0.569	0.696
Socratic Llama2 7B	3.42	0.162	0.216	0.576	0.670
Basic Llama2 7B	0.494	0.120	0.092	0.535	0.582
Random Llama2 7B	0.210	0.091	0.063	0.444	0.312





Example of conversation

Is replicability necessary in the production of knowledge?

Replicability is a crucial aspect of scientific research as it enables scientists to build upon previous findings and establish the reliability and validity of their results. Without replicability, research findings may be difficult to verify or validate, leading to a lack of confidence in the conclusions drawn from them...



Traditional chatbot



Example of conversation

Is replicability necessary in the production of knowledge?



Learner

Clarification

What do you mean by "replicability"?

Replicability means being able to reproduce or verify the results of an experiment or study by repeating it with similar methods and obtaining similar results.

Reasons

Why do you think you need replicability in order to produce knowledge?

Replicability is necessary for producing knowledge because it allows us to verify and confirm the accuracy of findings through independent repetition of experiments or observations.

Alternative viewpoints

How would you respond to someone who **argues that replication is not necessary** in this case because the sample size was so small?



Socratic
chatbot



Winning the critical thinking questions shared task: LLM-based question generation and selection



OBJECTIVE

The dataset Interventions of real debates, annotated with *Argumentation Schemes* and a set of *Critical Questions* and their usefulness for challenging the arguments:

Useful, *Unhelpful*, or *Invalid*

The task Generate three critical questions that are *Useful*



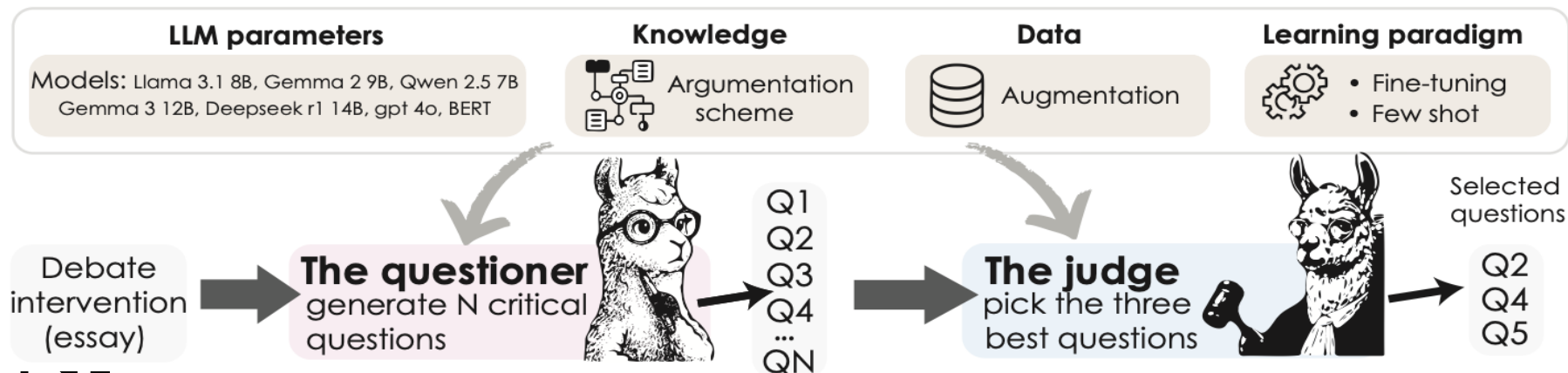
EVALUATION

- Compute **cosine similarity** of the question with the annotated references
- Limitations of automatic assessment** in such a complex and open task



METHOD

Questioner-Judge LLM architecture based on critical thinking theory: analytic, creative, and evaluative dimensions



RESULTS

Questioner	Scheme in the prompt	#Questions generated	Judge	% Useful	% Un-matched
Gemma 3 12B	without	8	-	40.5	46.0
Llama FT ***	without	8	-	56.6	29.4
Llama 3.1 8B	without	8	Deepseek 14B	57.4	27.0
Gemma 2 9B	with	8	Llama 3.1 8B	54.7	31.6
Llama 3.1 8B	with	6	Gemma 3 12B	49.4	34.1
Llama 3.1 8B	both	12	Gemma 2 9B	62.4	25.7
GPT 4o	both	12	GPT 4o	62.7	25.0

***Llama 3.1 8B fine-tuned

Best configuration

- Questioner + Judge
- With and without schemes
- Small open-source LLMs: Llama 3.1 8B, Gemma 2 9B
- Chat GPT 4o



Leveraging small LLMs for argument mining in education:

Argument component identification, classification, and Assessment



OBJECTIVE

Performing **argument mining** with **small open - source** LLMs:

- Component segmentation
- Type classification
- Quality assessment

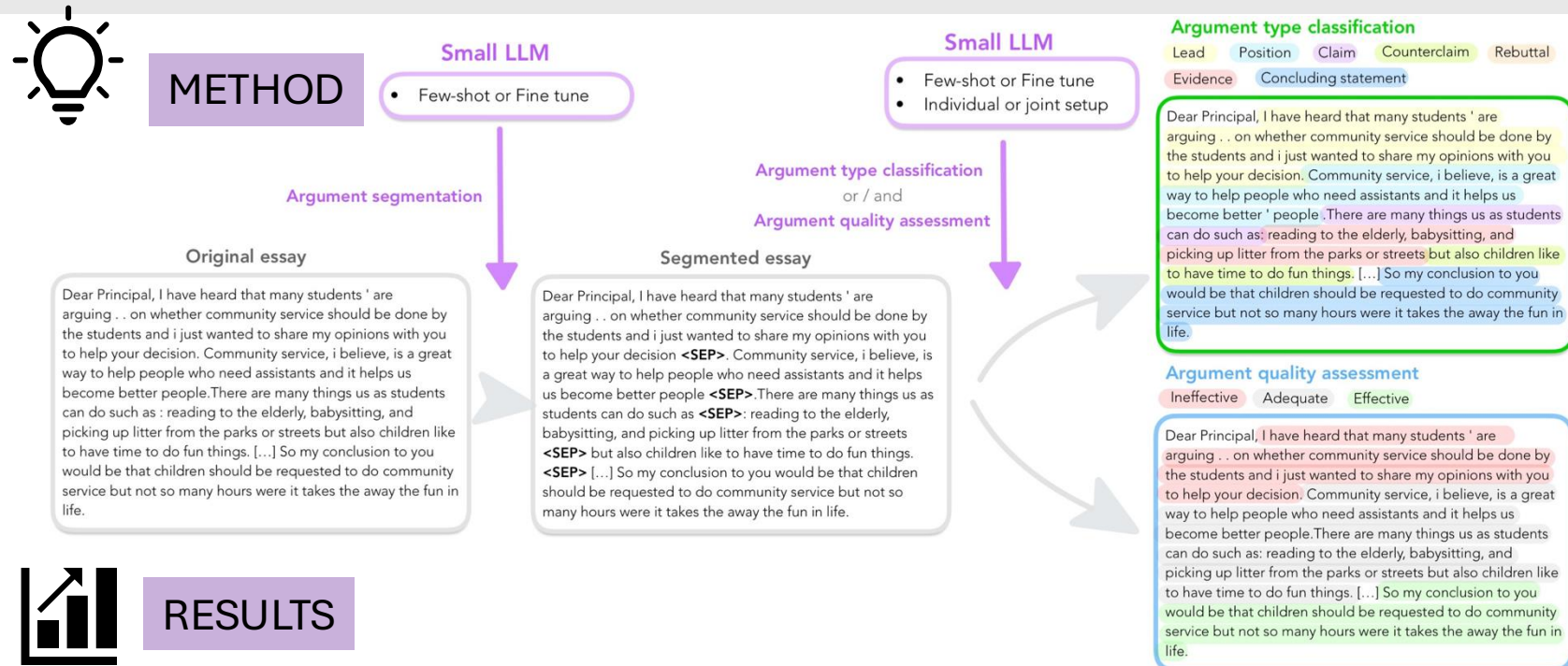


EVALUATION

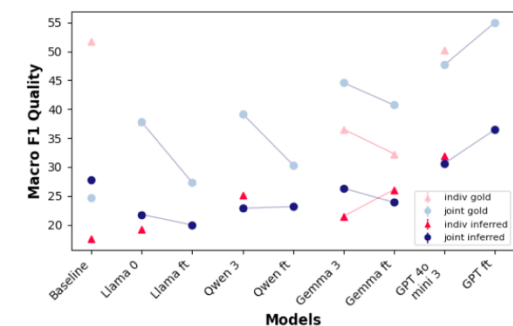
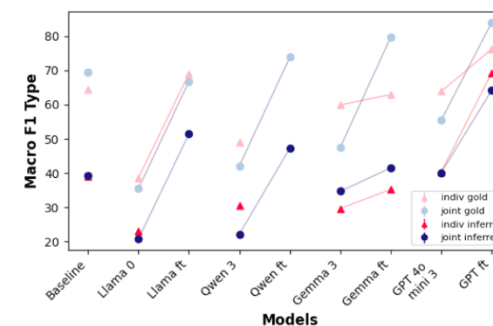
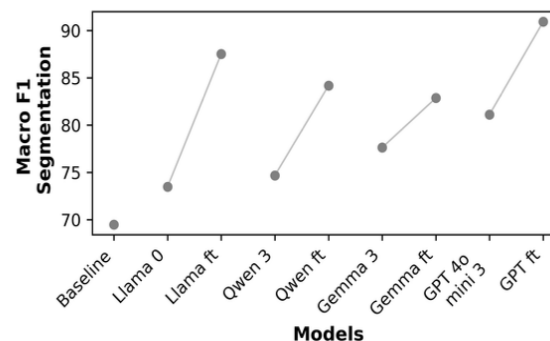
- *Dataset*: Feedback Prize
- *Metric*: Macro-averaged F1
- *Baseline*: BERT (two heads) and Longformer



METHOD



RESULTS



[5] Favero, Pérez-Ortiz, Käser, & Oliver. [Argument Mining in Education: Exploring the Potential of Open-source Small LLMs for Argument Classification and Assessment](#). *AI for Education Tools, Opportunities, and Risks in the Generative AI Era @ AAAI2025*

Thank you for your attention



- **Project presentation:** <https://ellisalicante.org/maike>
- **Posters:**
 - <https://web.ua.es/de/phdinf/documentos/jdi-2024/lucilealysfavero.pdf>
 - <https://web.ua.es/es/phdinf/documentos/jdi-2025/lucilealysfaveromontero.pdf>
- **Scientific papers:** <https://ellisalicante.org/people/lucilefavero-en/>
- **Talk:** <https://web.ua.es/en/iuii/noticias/talk-lucile-favero-towards-student-centric-ai-supported-learning-teaching-chatbots-to-ask-the-right-questions.html>
- **Contact:** lucile@ellisalicante.org <http://lucilefavero.github.io/>