Sufficient Context

A New Lens on RAG Systems

Cyrus Rashtchian

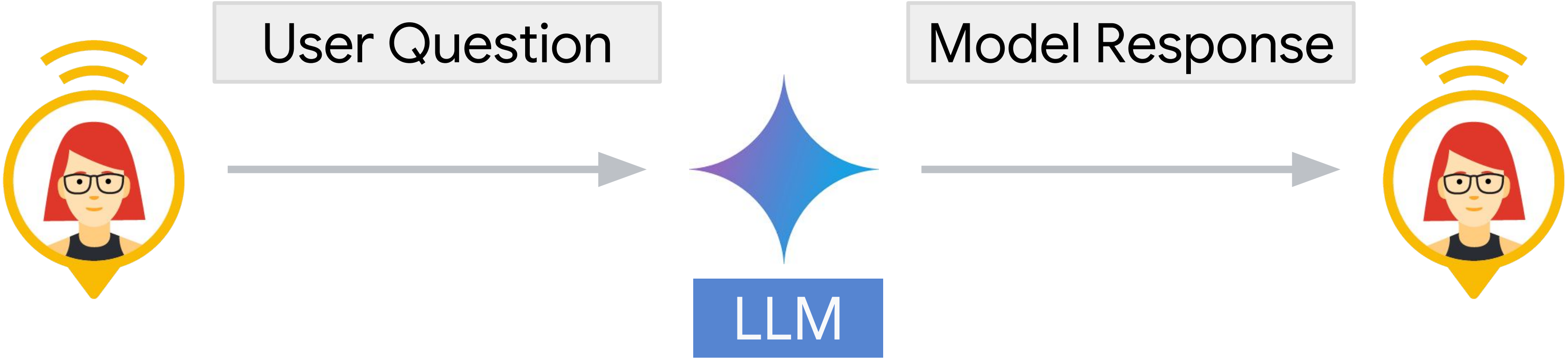
w/ Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly

To appear at ICLR 2025

February 10, 2025

Google Research

Retrieval Augmented Generation (RAG)

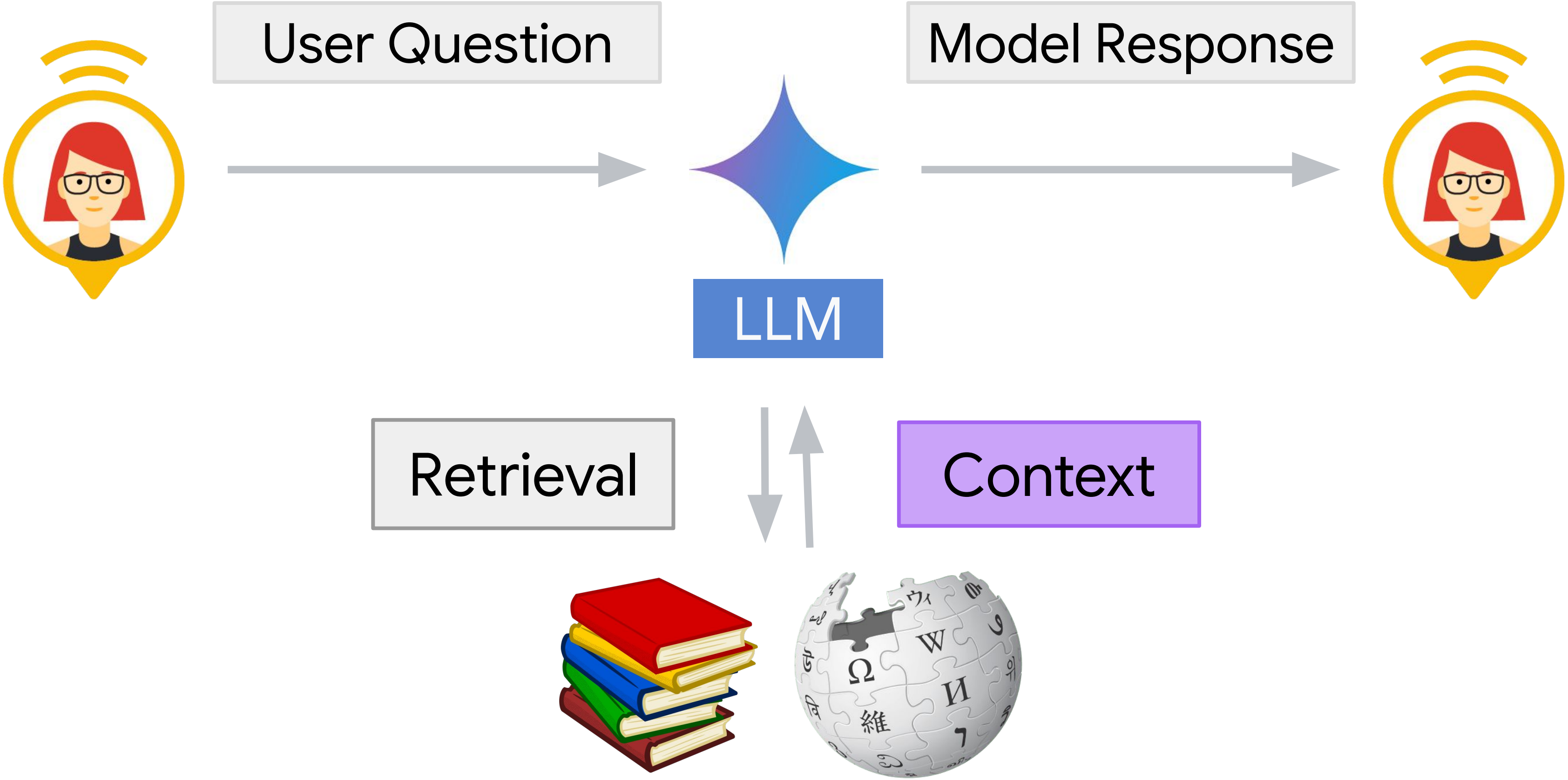


User Question

Model Response

LLM

Retrieval Augmented Generation (RAG)



User Question

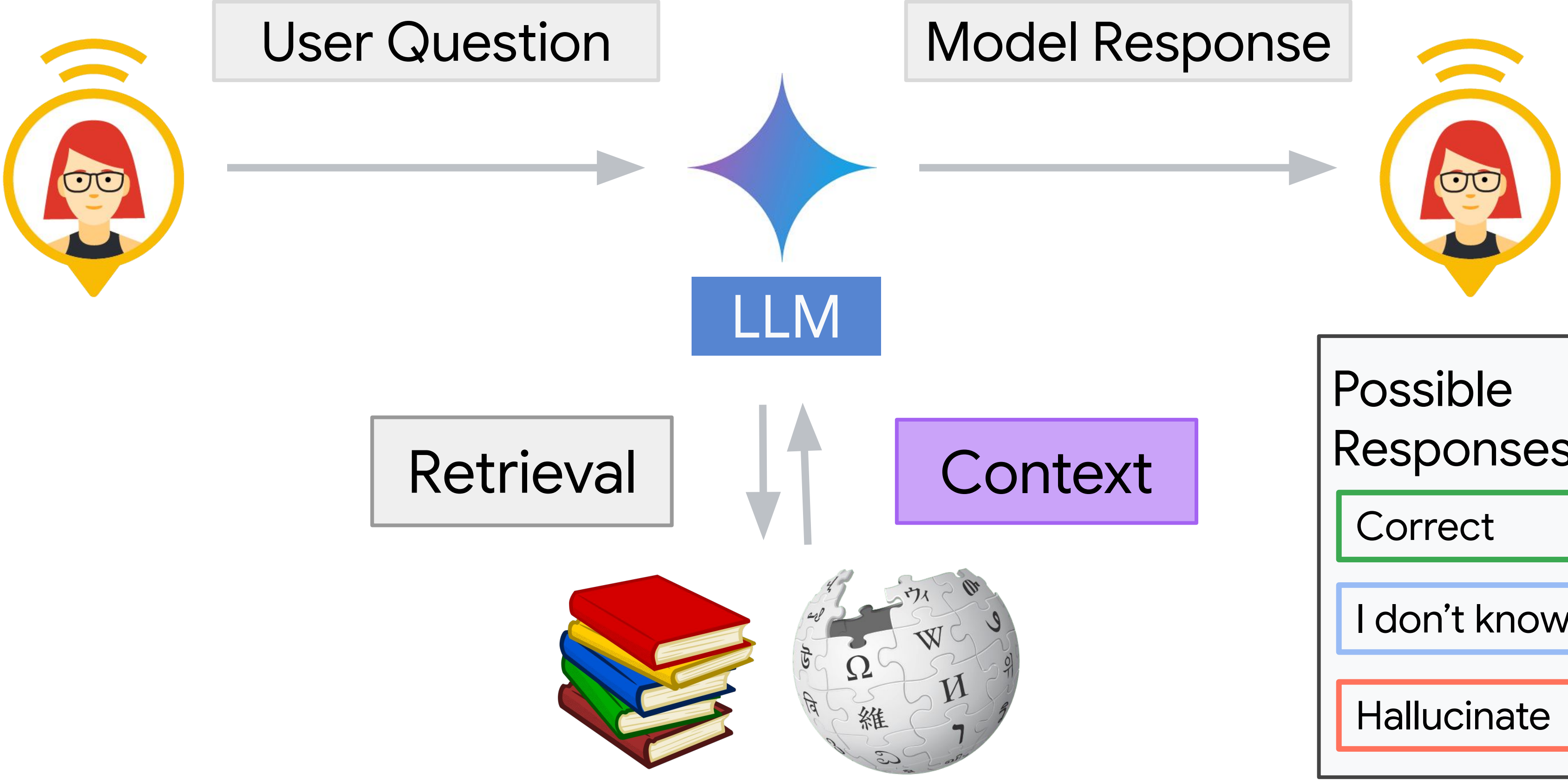
Model Response

LLM

Retrieval

Context

Retrieval Augmented Generation (RAG)



User Question

Model Response

LLM

Retrieval

Context

Possible Responses

- Correct
- I don’t know
- Hallucinate

Motivating Questions

RAG systems produce incorrect responses. What’s the cause?

- Does the retrieved context lack information?
- Do LLMs hallucinate when they should know the answer?
- How can we reduce the fraction of incorrect answers?

Main Contributions

1. Define the notion of sufficient context for RAG systems

2. Design a sufficient context AutoRater (achieving 93% accuracy)

3. Analyze SOTA LLMs
    a. Expect: Better with sufficient context (+20% to +40% correct)
    b. Surprising: correct 35–62% of the time with insufficient context

4. Construct a new selective generation method
    a. Improves correct answers (among total model responses) by 2–10% for Gemini 1.5 Pro, GPT 4o, and Gemma 2 27B

Outline

01 Sufficient Context Definition

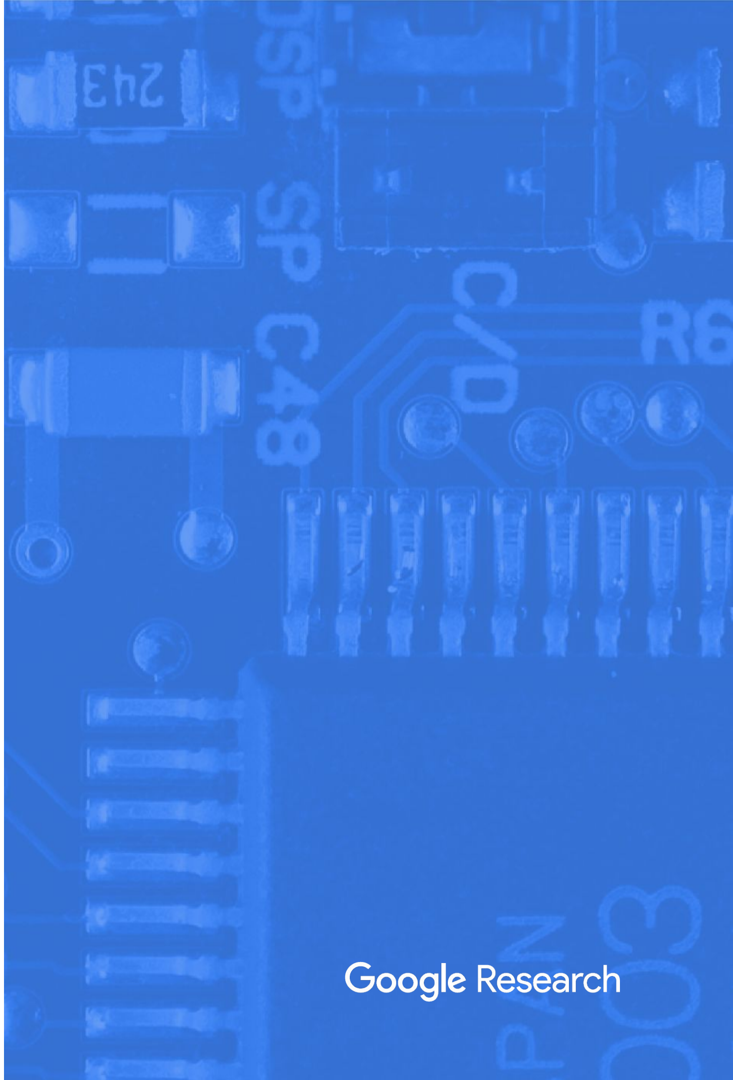
02 AutoRater Development

03 Dataset and Model Analysis

04 Intervention: Selective Generation

Sufficient Context

Definition & Examples

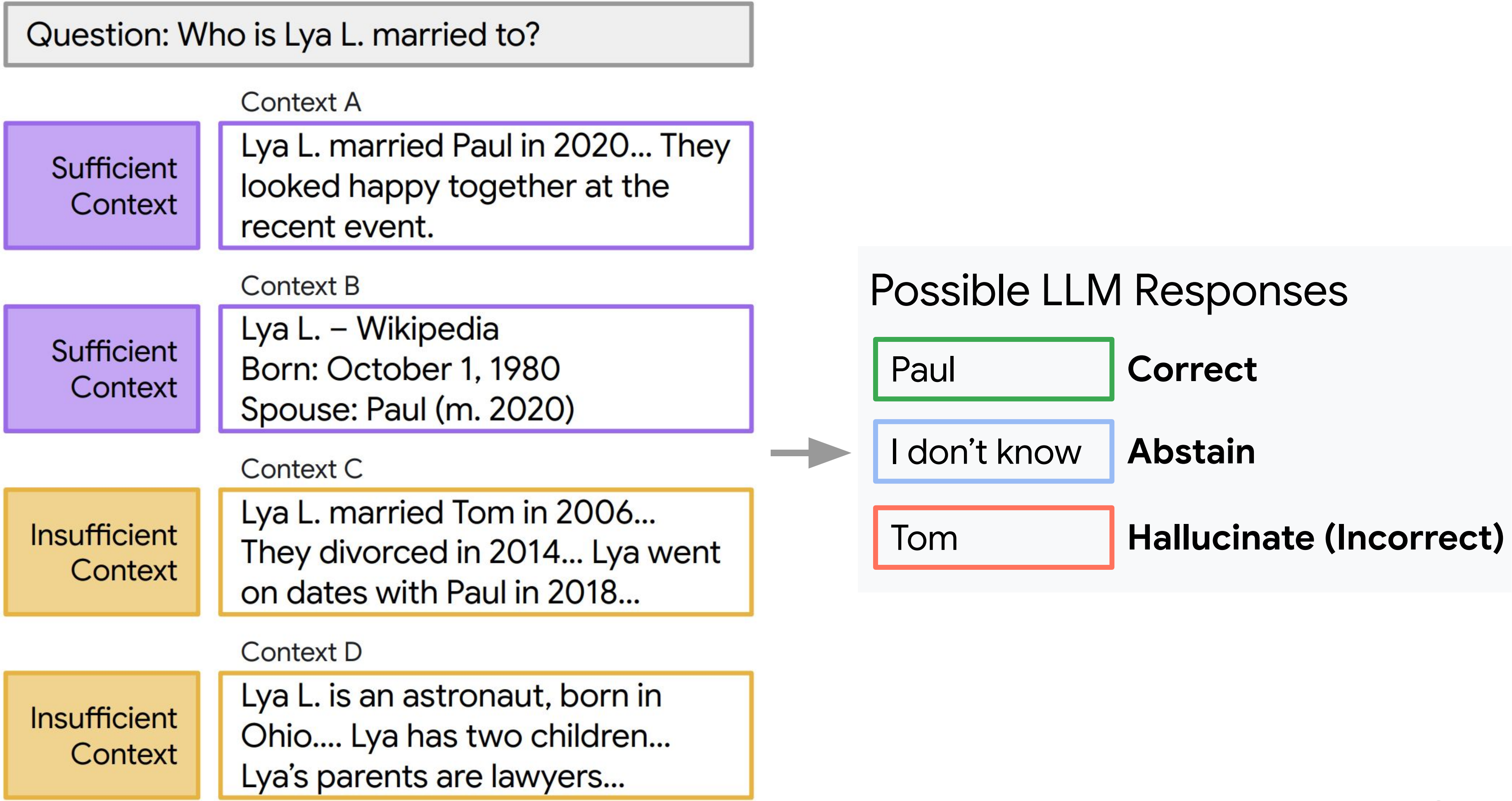


Sufficient Context Definition

Evaluate query & sources (no model response, no ground truth)

Sufficient Context. The context is sufficient to answer a question if it contains all the necessary information to provide a definitive answer to the question.

Insufficient Context. Otherwise, a context is insufficient. A context may also be insufficient if the question requires specialized knowledge that is not provided in the context or if the information is incomplete, inconclusive, or contradictory.



Question: Who is Lya L. married to?

Context A

Sufficient Context

Lya L. married Paul in 2020... They looked happy together at the recent event.

Context B

Sufficient Context

Lya L. – Wikipedia
Born: October 1, 1980
Spouse: Paul (m. 2020)

Context C

Insufficient Context

Lya L. married Tom in 2006... They divorced in 2014... Lya went on dates with Paul in 2018...

Context D

Insufficient Context

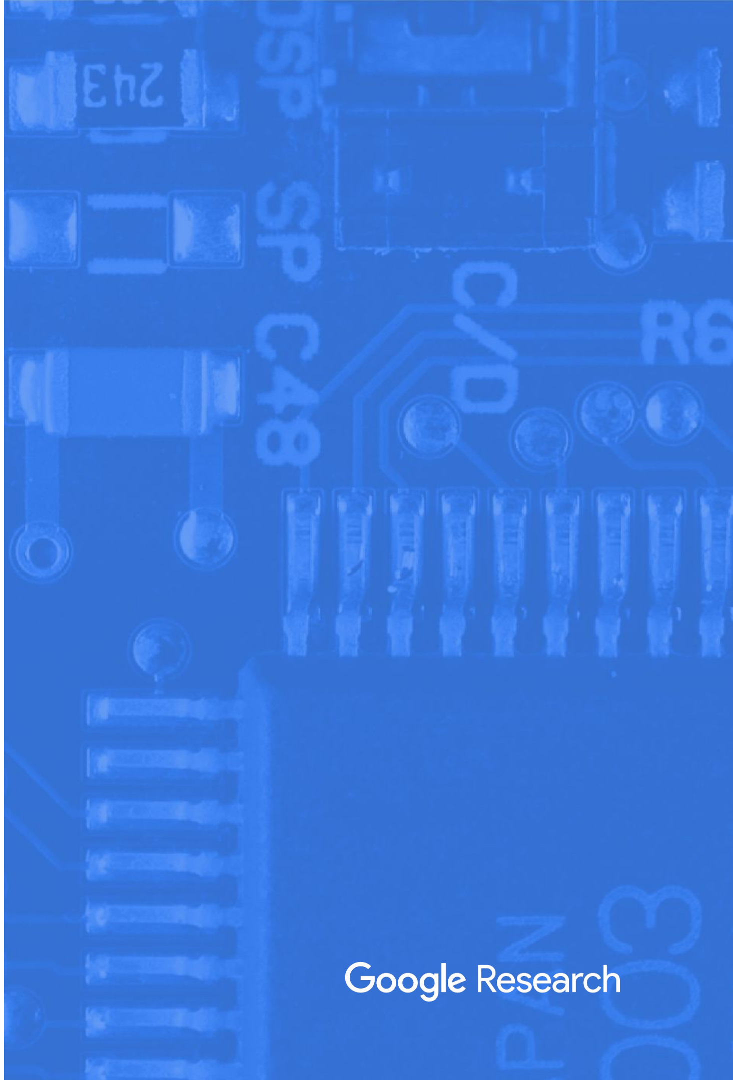
Lya L. is an astronaut, born in Ohio.... Lya has two children... Lya's parents are lawyers...

Possible LLM Responses

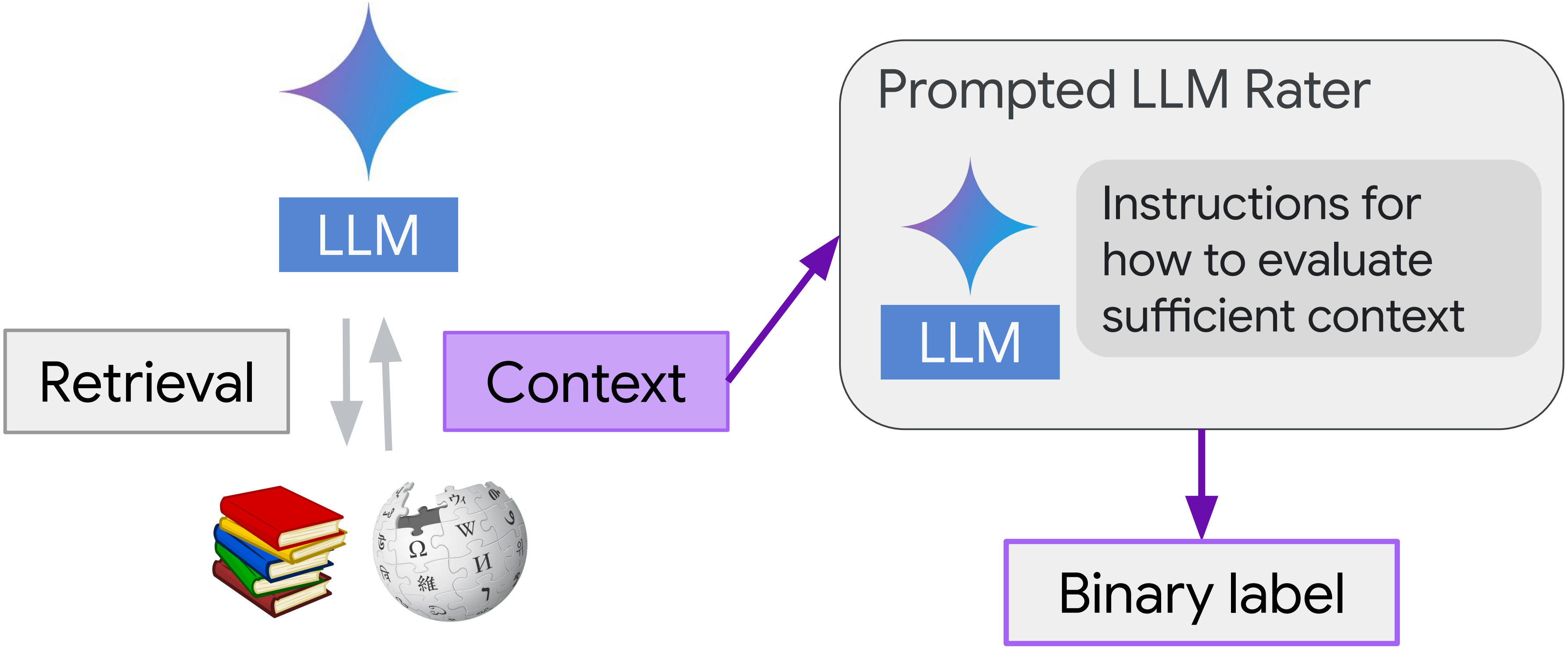
Paul	**Correct**
I don't know	**Abstain**
Tom	**Hallucinate (Incorrect)**

AutoRater

Develop & Evaluate



AutoRater for Sufficient Context



LLM

Retrieval

Context

Prompted LLM Rater

LLM

Instructions for how to evaluate sufficient context

Binary label

AutoRater for Sufficient Context

User

Input Prompt

You are a helpful assistant.... <preamble for sufficient context autorater>

Question: How many nobel peace prize laureates have been assassinated?

Context: <long context ~500k characters, list of laureates and info about them>

Is the context above sufficient to answer the question above? Please answer TRUE or FALSE and also provide a reason for your response.

✦ Assistant

Model Response

TRUE

Reason: The context provides a list of Nobel Peace Prize laureates and additional information about their assassinations. By examining both, we can identify the laureates who were assassinated (Martin Luther King Jr., Anwar El-Sadat, and Yitzhak Rabin) and count them, directly answering the question.

AutoRater for Sufficient Context

Can we automatically determine sufficient context?

Dataset: we labeled 115 instances, sampled from public benchmarks

Best model

- Prompted Gemini 1.5 Pro with 1-shot example (93% Accuracy)

Baselines, without ground truth (GT) answer

- Prompted Gemini 1.5 Pro (0-shot)
- FLAMe 24B: foundational AutoRater model

Baselines, require GT answer

- AutoAIS: entailment of the GT answer
- Simple check for “contains GT answer”

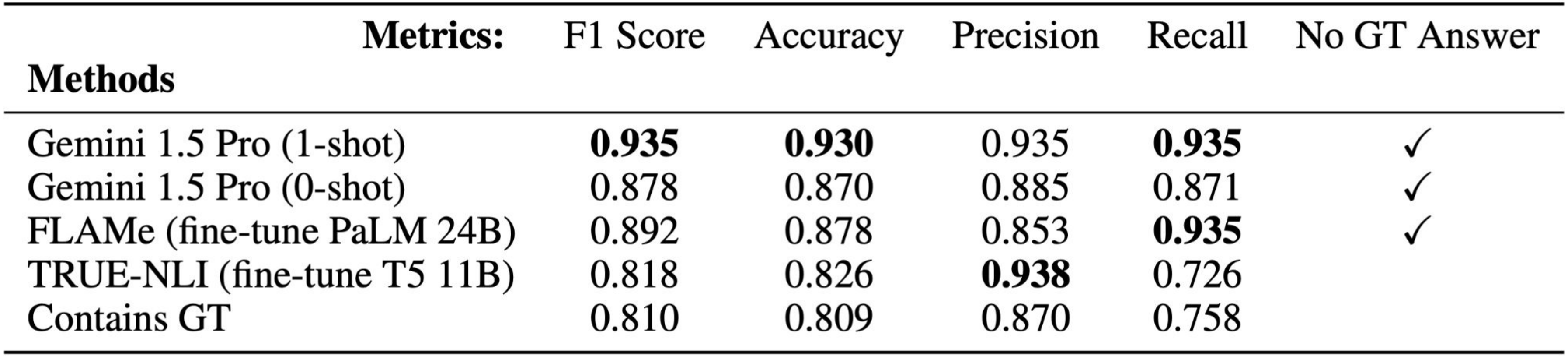
AutoRater for Sufficient Context

Can we automatically determine sufficient context?

Dataset: we labeled 115 instances, sampled from public benchmarks

Best model

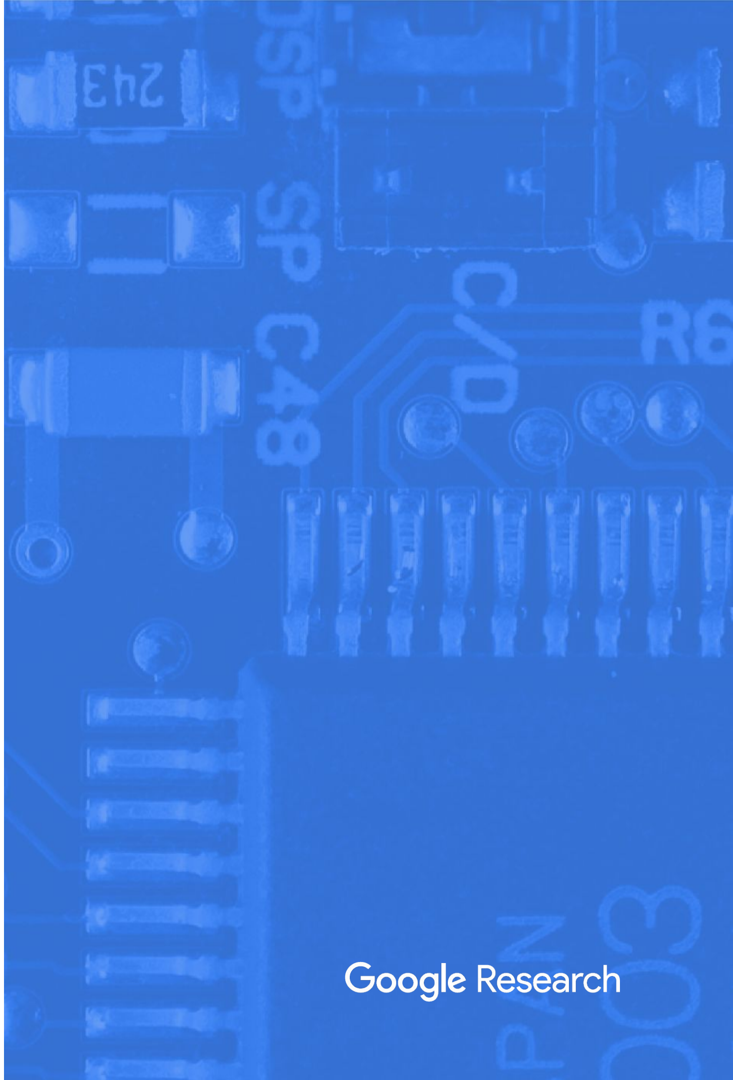
- Prompted Gemini 1.5 Pro with 1-shot example (93% Accuracy)



Metrics: **Methods**	F1 Score	Accuracy	Precision	Recall	No GT Answer
Gemini 1.5 Pro (1-shot)	**0.935**	**0.930**	0.935	**0.935**	✓
Gemini 1.5 Pro (0-shot)	0.878	0.870	0.885	0.871	✓
FLAMe (fine-tune PaLM 24B)	0.892	0.878	0.853	**0.935**	✓
TRUE-NLI (fine-tune T5 11B)	0.818	0.826	**0.938**	0.726	
Contains GT	0.810	0.809	0.870	0.758	

Analysis

Datasets & Models



Google Research

Analysis

Using our Sufficient Context AutoRater (Gemini 1.5 Pro, 1-shot)

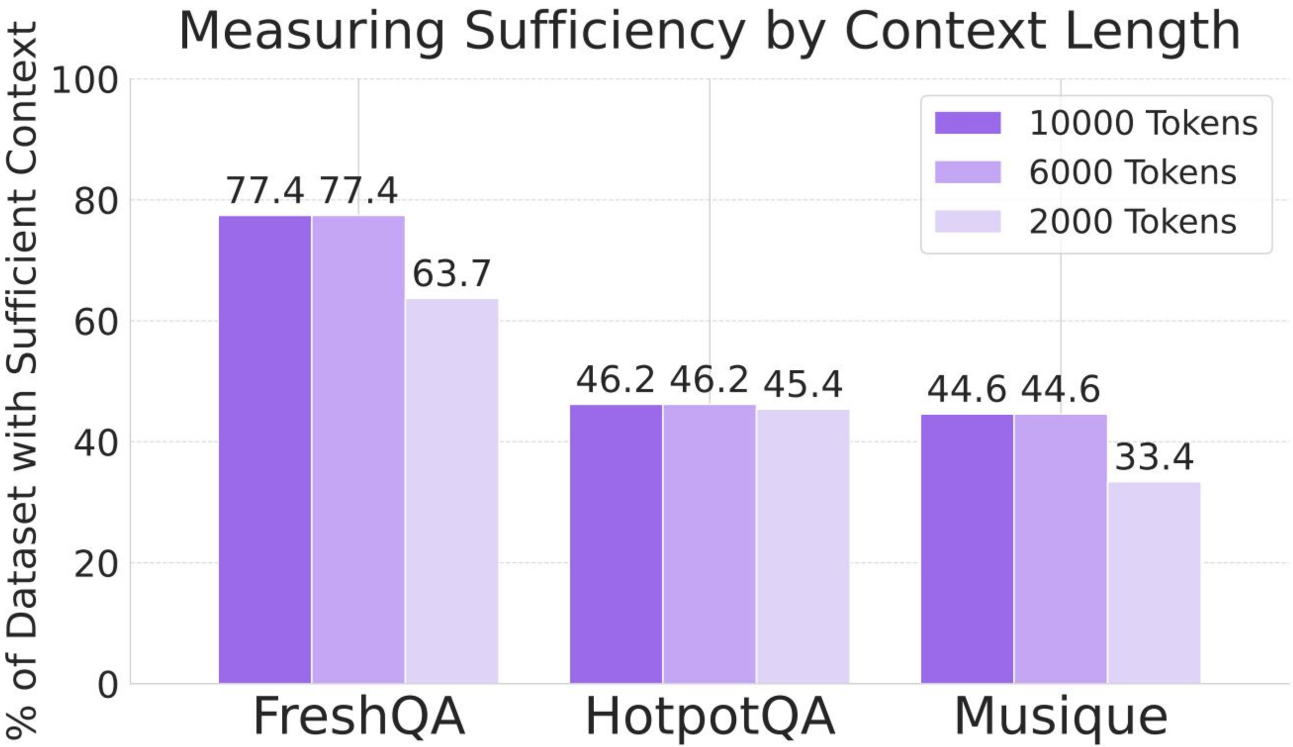
- **Scalably label instances** in datasets as sufficient or insufficient
- **Measure % of dataset with sufficient context** (unexpectedly low)
- **Categorize model performance** w/ sufficient vs. insufficient context

Datasets:

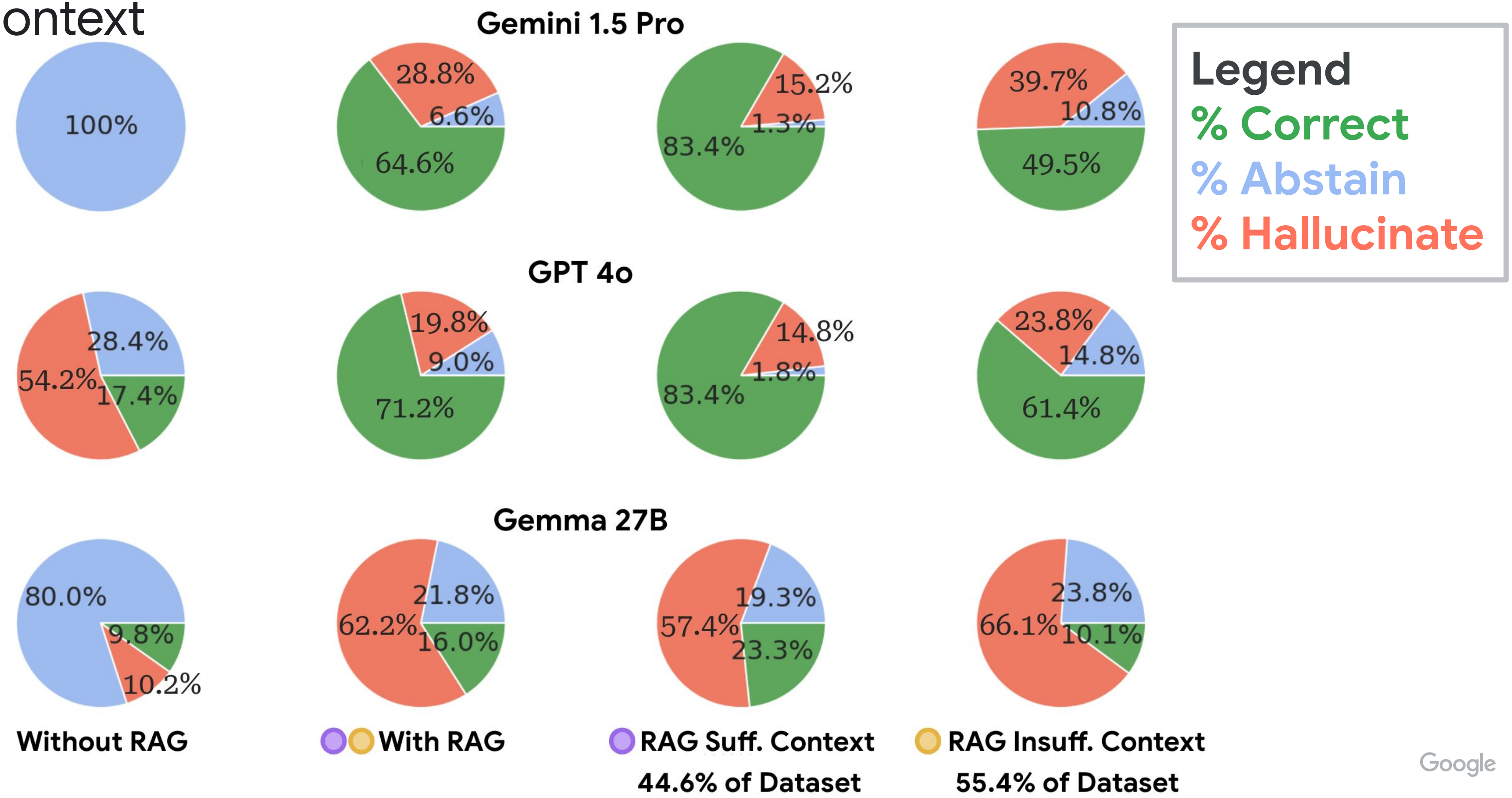
- FreshQA: time-sensitive questions
- Musique: multi-hop questions
- HotPotQA: single- and multi-hop questions

Dataset Analysis

Compare % of instances with Sufficient Context vs. Length of Context

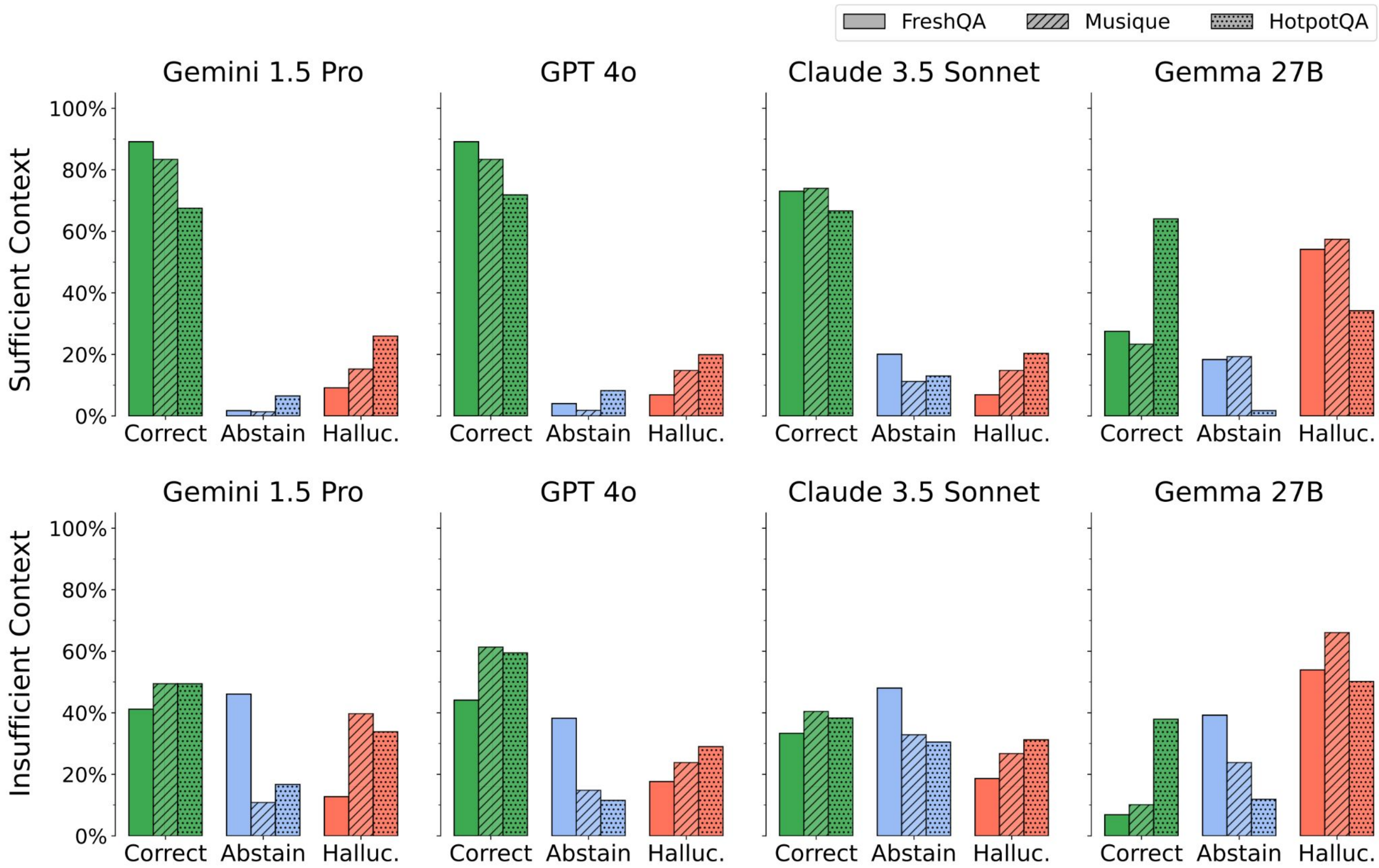


Models Hallucinate More with RAG, Especially w/ Insufficient Context



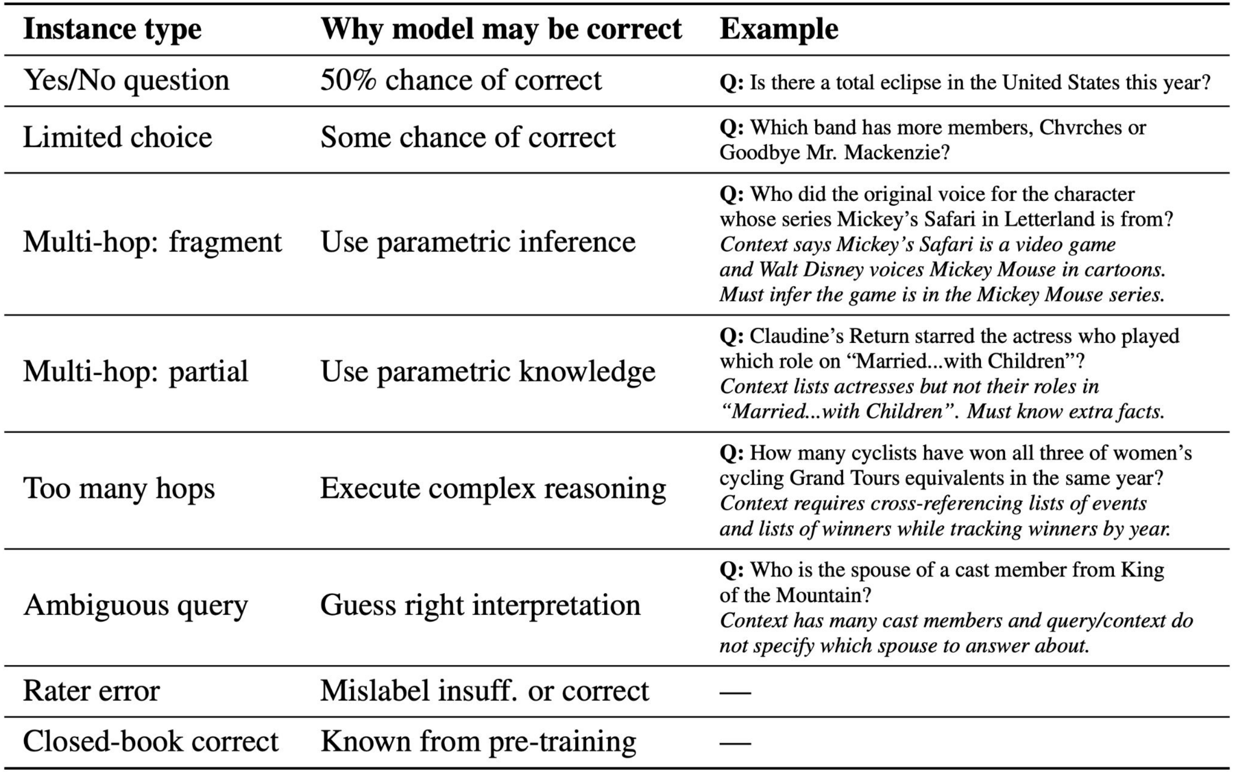
Model Analysis: 3 Datasets

% Correct, **% Abstain**, **% Hallucinate**



Insufficient Context and Model is Correct

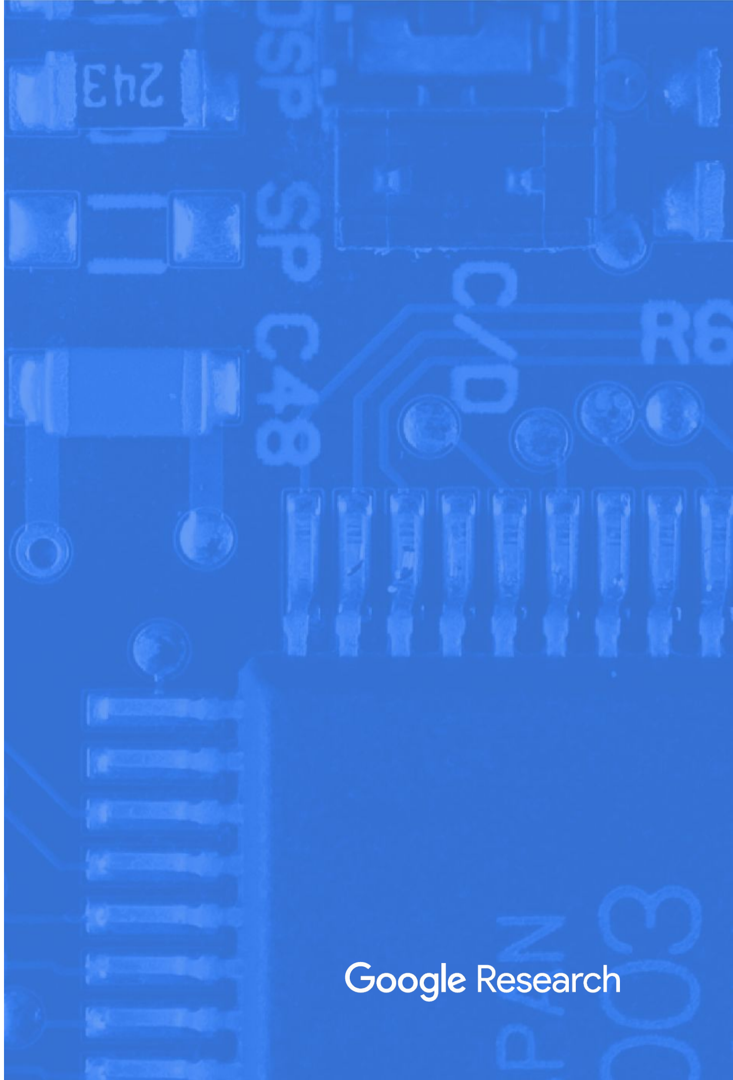
Why are models correct on 35–62% of instances w/ insufficient context?



Instance type	**Why model may be correct**	**Example**
Yes/No question	50% chance of correct	**Q:** Is there a total eclipse in the United States this year?
Limited choice	Some chance of correct	**Q:** Which band has more members, Chvrches or Goodbye Mr. Mackenzie?
Multi-hop: fragment	Use parametric inference	**Q:** Who did the original voice for the character whose series Mickey's Safari in Letterland is from? *Context says Mickey's Safari is a video game and Walt Disney voices Mickey Mouse in cartoons. Must infer the game is in the Mickey Mouse series.*
Multi-hop: partial	Use parametric knowledge	**Q:** Claudine's Return starred the actress who played which role on "Married...with Children"? *Context lists actresses but not their roles in "Married...with Children". Must know extra facts.*
Too many hops	Execute complex reasoning	**Q:** How many cyclists have won all three of women's cycling Grand Tours equivalents in the same year? *Context requires cross-referencing lists of events and lists of winners while tracking winners by year.*
Ambiguous query	Guess right interpretation	**Q:** Who is the spouse of a cast member from King of the Mountain? *Context has many cast members and query/context do not specify which spouse to answer about.*
Rater error	Mislabel insuff. or correct	—
Closed-book correct	Known from pre-training	—

Intervention

Selective Generation



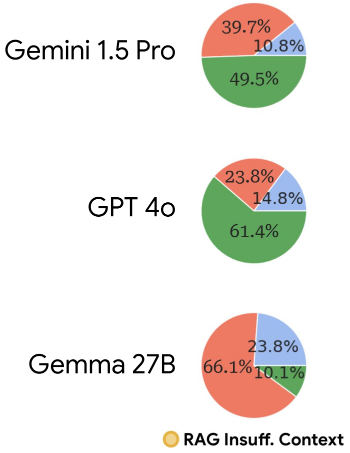
Reducing Hallucinations

Models hallucinate more w/ RAG, both sufficient and insufficient context

How do we fix this?

- Fine-tuning does not really work (seems hard to get models to abstain)
- Only answering with sufficient context misses out on a lot of correctness

Model performance with insufficient context



Always abstaining when the context is insufficient would leave a lot of correctness on the table (saying “I don’t know” too much)

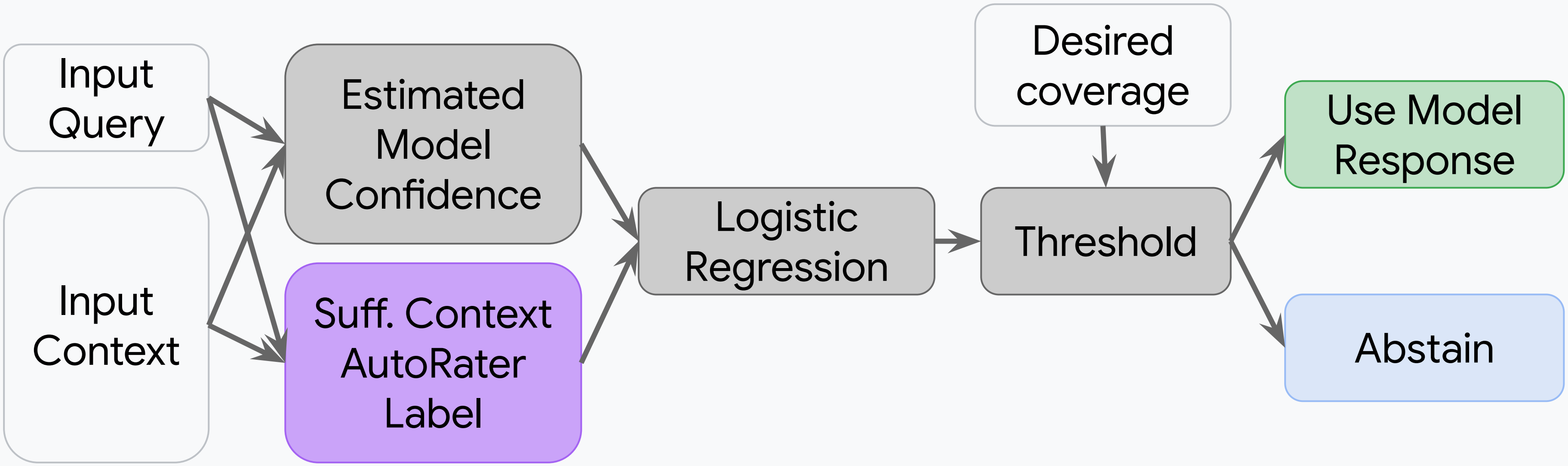
Our Approach: Selective Generation

- Train small model to decide to answer or abstain
    - Use logistic regression with model confidence **and** sufficient context label
- Choose threshold to balance coverage and selective accuracy

Self-reported model confidence

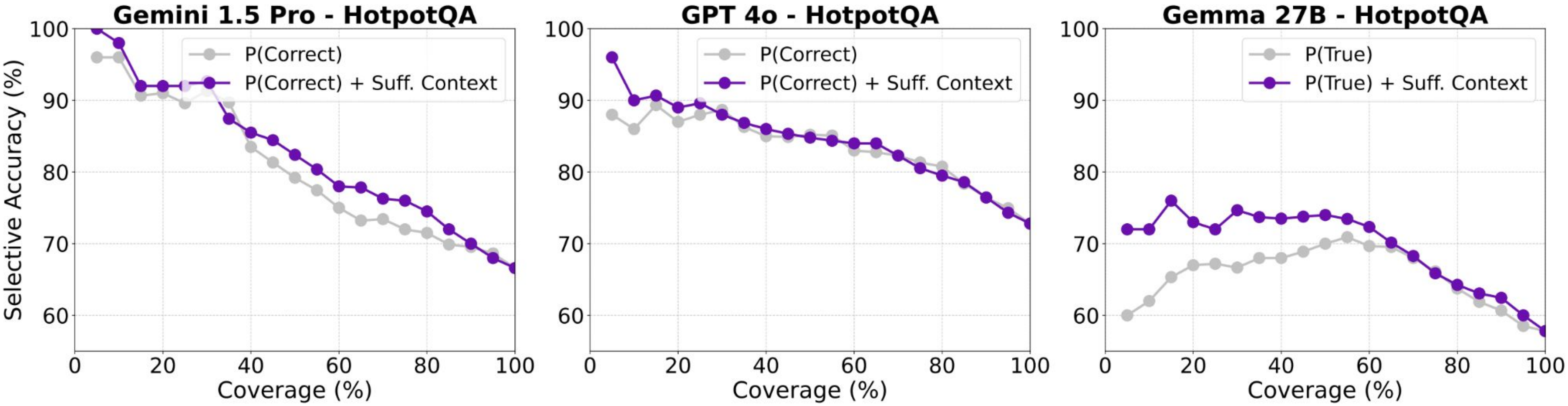
- Small models like Gemma, we use P(True)
- Large models like Gemini/GPT, model estimates its confidence, which we call P(Correct)

Selective Generation Framework



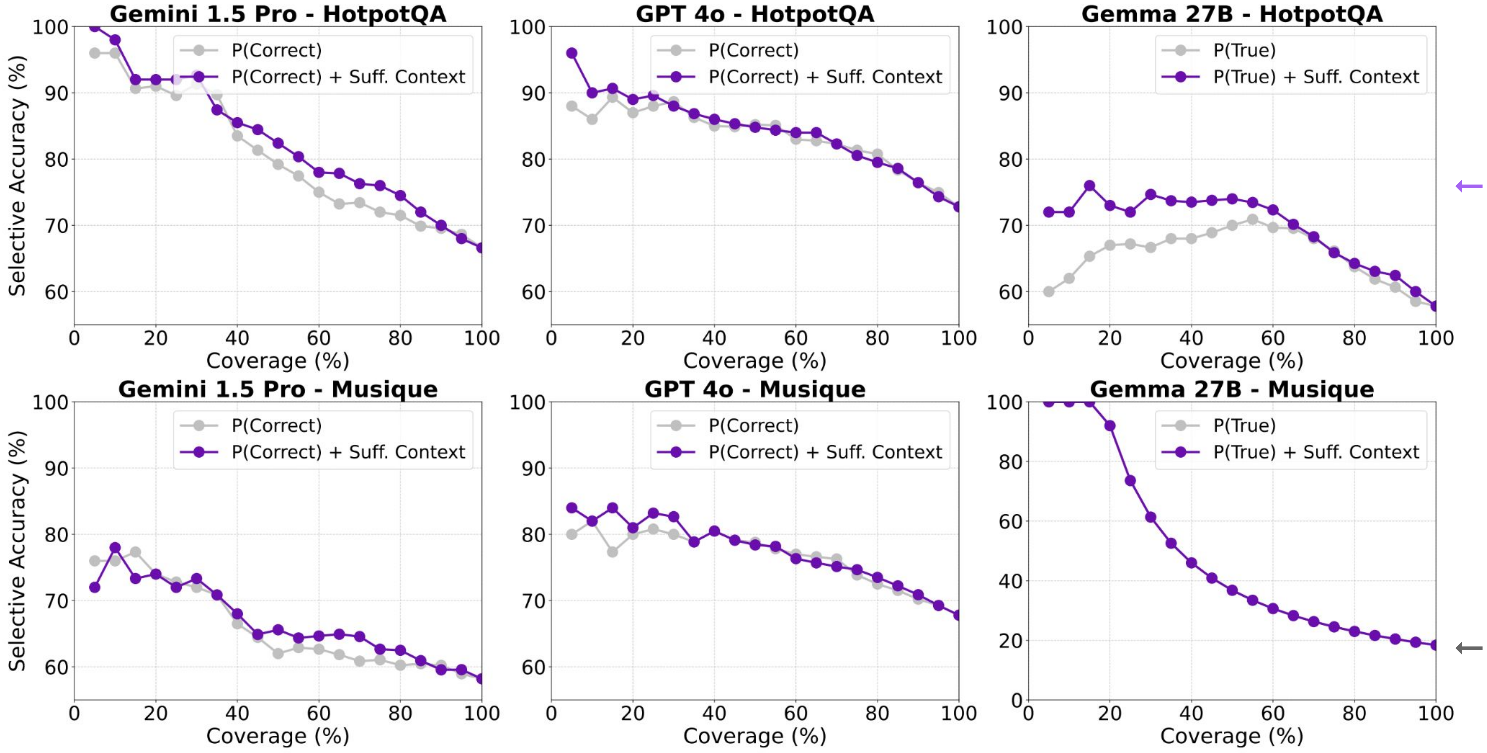
Selective Generation vs. Abstention

Choose threshold θ and only respond when f(confidence, suff. context) > θ



Selective Generation vs. Abstention

Choose threshold θ and only respond when f(confidence, suff. context) > θ



Improvement! ⟶

Improvement! ⟶

⟵ Very Big Improvement!

⟵ Coeff. = 0 for Suff. Context

Google

Conclusion

1. Defined the notion of sufficient context for RAG systems
2. Designed a sufficient context AutoRater (93% accuracy)
3. Analyzed SOTA LLMs (Gemini, GPT 4, Claude, Gemma, Llama)
4. Constructed a new selective generation method

Open Questions

1. Extend sufficient context to multi-modal (e.g., PDFs, images)

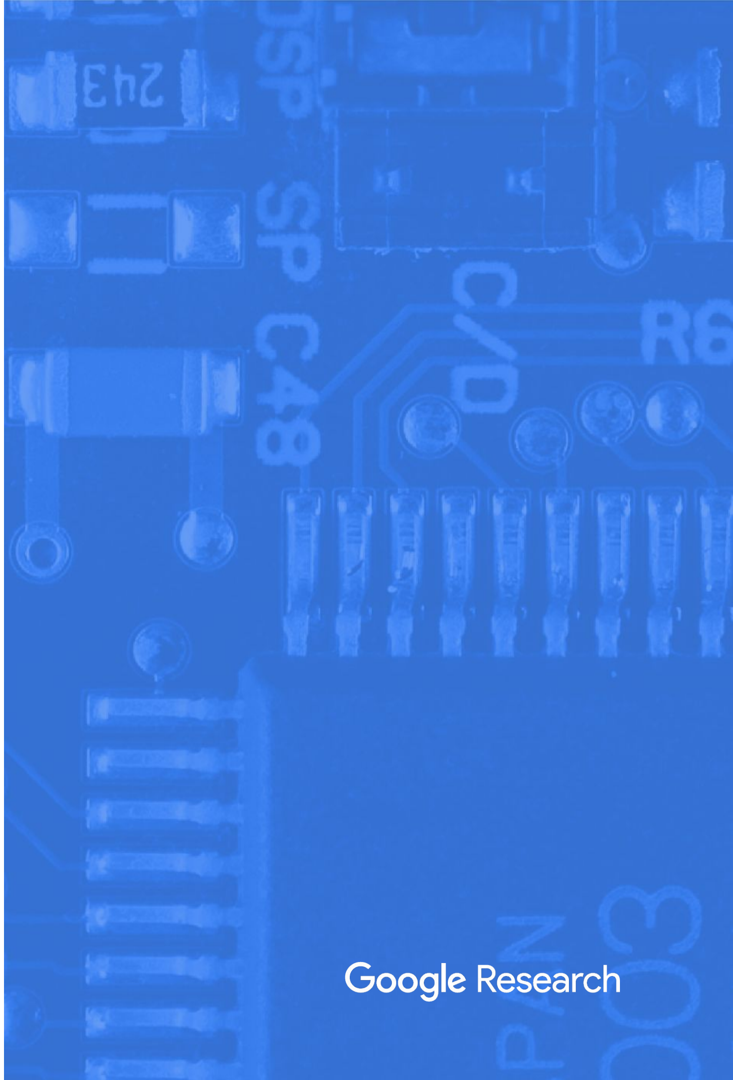
2. Develop autoraters for other “comprehension” tasks
   → Agents can be a powerful tool for data analysis

3. Create small LLMs that have high accuracy w/ sufficient context
   → Should be feasible, just synthesizing info from retrieval

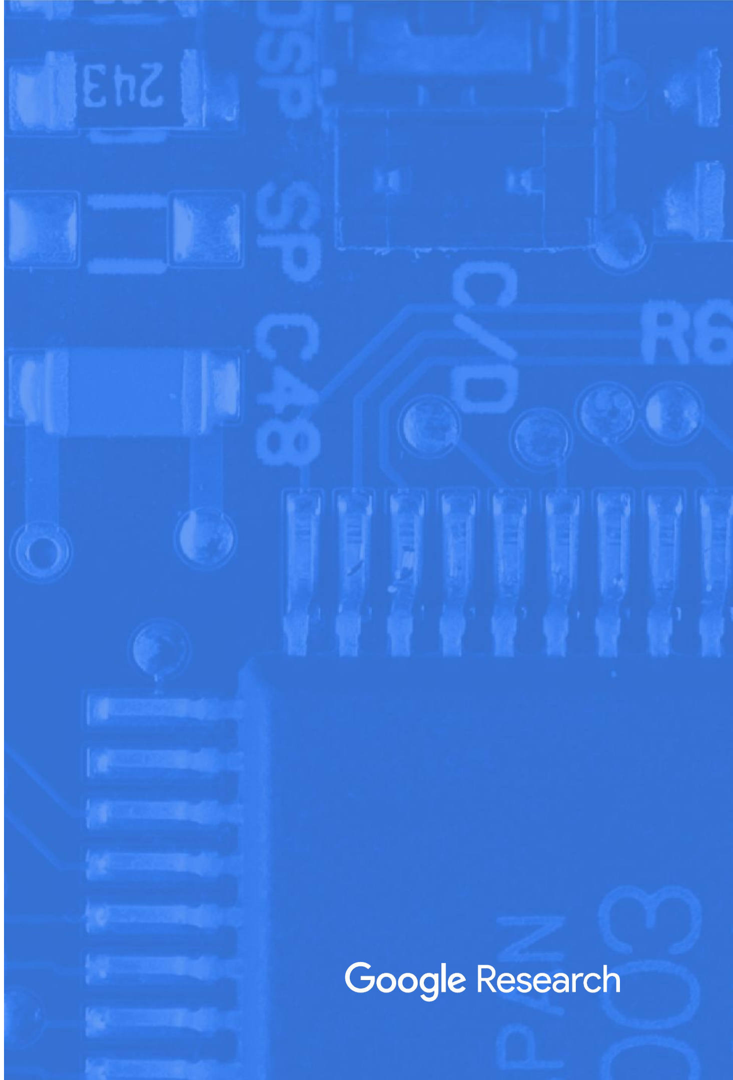
4. Improve LLMs so they abstain instead of hallucinate
   → Better RL / Fine-tuning?

Thanks!

Cyrus Rashtchian
cyroid@google.com



Appendix



Fine-tuning to Encourage Abstention

Experiment: what if we change some answers to “I don’t know” before fine-tuning? Does that help?

- First, we sample 2000 instances
- **Data Mix 1:** fine-tune on these instances, keep their ground truth answer
- **Data Mix 2:** choose 400 examples (20%) at random, change answer to “I don’t know” before fine-tuning.
- **Data Mix 3:** instead randomly choose 400 examples (20%) that our autorater labels as insufficient context and change their answer to “I don’t know” while keeping the other answers as the ground truth.

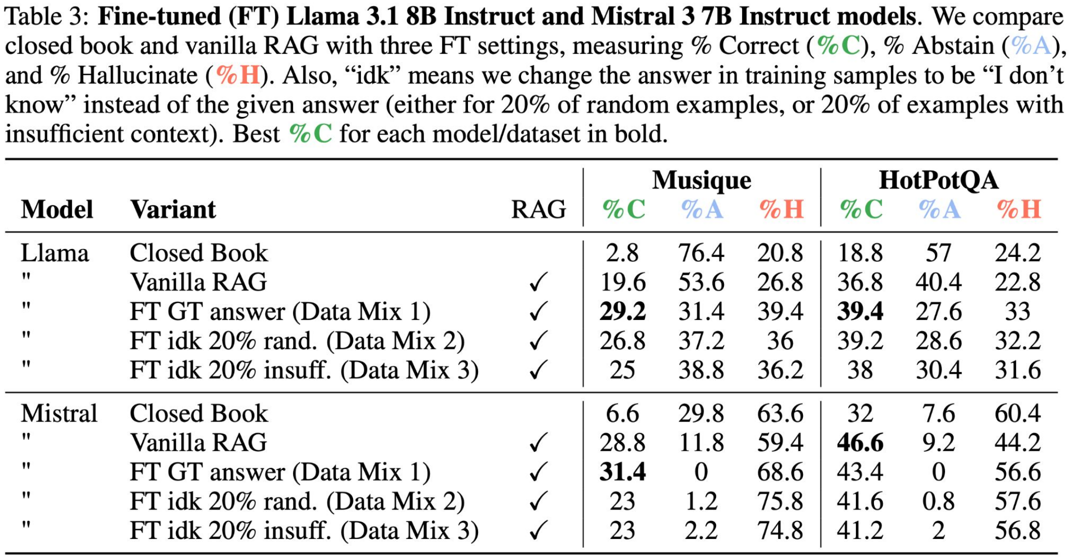


Table 3: **Fine-tuned (FT) Llama 3.1 8B Instruct and Mistral 3 7B Instruct models**. We compare closed book and vanilla RAG with three FT settings, measuring % Correct (%C), % Abstain (%A), and % Hallucinate (%H). Also, “idk” means we change the answer in training samples to be “I don’t know” instead of the given answer (either for 20% of random examples, or 20% of examples with insufficient context). Best %C for each model/dataset in bold.

			Musique			HotPotQA		
Model	**Variant**	RAG	**%C**	**%A**	**%H**	**%C**	**%A**	**%H**
Llama	Closed Book		2.8	76.4	20.8	18.8	57	24.2
"	Vanilla RAG	✓	19.6	53.6	26.8	36.8	40.4	22.8
"	FT GT answer (Data Mix 1)	✓	**29.2**	31.4	39.4	**39.4**	27.6	33
"	FT idk 20% rand. (Data Mix 2)	✓	26.8	37.2	36	39.2	28.6	32.2
"	FT idk 20% insuff. (Data Mix 3)	✓	25	38.8	36.2	38	30.4	31.6
Mistral	Closed Book		6.6	29.8	63.6	32	7.6	60.4
"	Vanilla RAG	✓	28.8	11.8	59.4	**46.6**	9.2	44.2
"	FT GT answer (Data Mix 1)	✓	**31.4**	0	68.6	43.4	0	56.6
"	FT idk 20% rand. (Data Mix 2)	✓	23	1.2	75.8	41.6	0.8	57.6
"	FT idk 20% insuff. (Data Mix 3)	✓	23	2.2	74.8	41.2	2	56.8

Results:
% Correct mostly goes up
% Abstain goes down vs. Vanilla RAG

Summary:
Not a good solution (needs better ideas)

Colors

text **Sufficient, dark** #A463F2

text **Sufficient, light** #CAA3F9

text **Sufficient, very light** #E3D1FB

text **Insufficient, dark** #EFB118

text **Insufficient, light** #FAE8C2

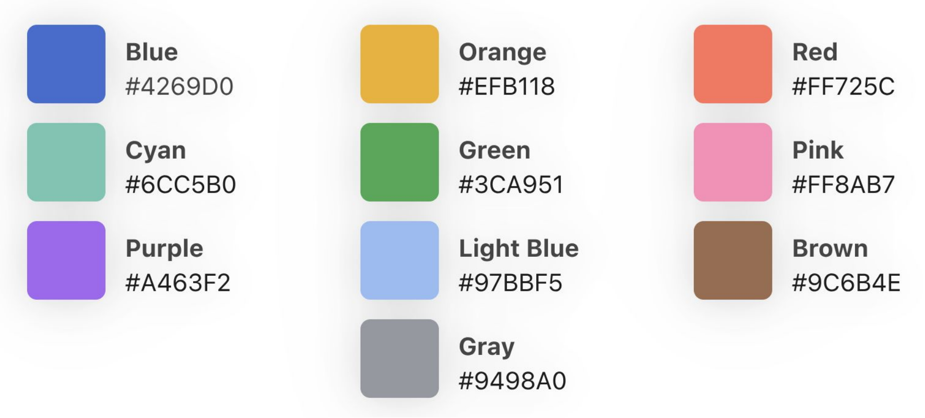
text **Insufficient, very light** #F5D283

text **Correct #3CA951**

text **Abstain #97BBF5**

text **Hallucinate #FF725C**

Inspiration = Observable 10



Blue #4269D0

Cyan #6CC5B0

Purple #A463F2

Orange #EFB118

Green #3CA951

Light Blue #97BBF5

Gray #9498A0

Red #FF725C

Pink #FF8AB7

Brown #9C6B4E