Precise **I**nformation **C**ontrol in Long-Form Text Generation

Jacqueline He · **Howard Yen** · **Margaret Li** · **Stella Li** · **Zhiyuan Zeng** · **Weijia Shi**

Yulia Tsvetkov · **Danqi Chen** · **Pang Wei Koh** · **Luke Zettlemoyer**

UWNLP

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

PLI

Ai2

0. *PIC* as a New Formulation

What exactly is Precise Information Control?

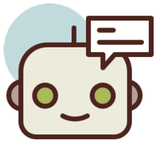
Research Question:

Can LMs produce long-form responses grounded only on given claims, without adding any unsupported ones?

Verifiable Claims

- PIC operates at the granularity of **verifiable claims** —short information units that can be unambiguously validated [Song et al., 2025].

Original Text:
Beetroot gets its beautiful red color from a special helper called betacyanin. Betacyanins are a type of anthocyanin, which are water-soluble pigments that commonly found in various fruits and vegetables... Betacyanin is like a superhero cape that makes the beetroot stand out!



LLM Claim Extractor

Verifiable Claims:
Claim 1: Beetroot gets its red color from betacyanin.
Claim 2: Betacyanins are a type of anthocyanin.
Claim 3: Anthocyanins are water-soluble pigments.
Claim 4: Anthocyanins are commonly found in various fruits and vegetables.

Song et al., 2025. VeriScore: Evaluating the factuality of verifiable claims in long-form text generation.

Verifiable Claims in **PIC**

- **Task:** Given an instruction and a list of verifiable claims, output a response that incorporates only these claims.

- Evaluation is exact and straightforward!

- We leverage LLM-based claim extractor and verification tools from prior work [Song et al., 2025], which correlate well with human judgment.

Verifiable Claims in **PIC**

- For comprehensiveness, we introduce two settings (based on user instruction):
  - **Full PIC:** The LM must include exactly all given claims in its response.
    - The ***F1*** between input and response claims is important.
  - **Partial PIC:** The LM chooses any relevant subset of claims to include in its response.
    - The ***precision*** between input and response claims is important.

Is that it?! This task seems very easy....

Is that it?! This task seems very easy....

🚨 **Spoiler alert:** We show that even SOTA LMs struggle against user control in this simple setting!

1. PIC-Bench

How can we evaluate PIC on today's LMs?

A New Benchmark

From SFT data

Input *I*: Give me a short biography on Elena Ferrante.

Output *y*: Elena Ferrante is a pseudonymous Italian novelist. Born in 1943, she is most famous for writing the Neapolitan Novels...

We can recast long-form generation samples from existing instruction-tuning datasets for PIC evaluation!

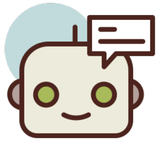
A New Benchmark

From SFT data

Input *I*: Give me a short biography on Elena Ferrante.

Output *y*: Elena Ferrante is a pseudonymous Italian novelist. Born in 1943, she is most famous for writing the Neapolitan Novels...

LLM Claim Extractor



Verifiable Claims *C*:

1. Elena Ferrante is a pseudonym.
2. Elena Ferrante is Italian.
3. Elena Ferrante is a novelist...

PIC task: given I and C, generate a response which can be decomposed into claims C'.

We can recast long-form generation samples from existing instruction-tuning datasets for PIC evaluation!

A New Benchmark

We instantiate PIC on **8 long-form generation tasks: 6 in the full setting and 2 in the partial setting.**

Task Name	PIC Type	N	$\bar{C}$	Example Instruction $\mathcal{I}$
EntityBios$_{PIC}$	Full	183	50.5	Generate a factual biography about Suthida.
PopBios-P$_{PIC}$	Full	111	20.1	Give me a biography on Erwin Schrödinger, the scientist who discovered Quantum Mechanics, Schrödinger's Cat Thought Experiment.
PopBios-CF$_{PIC}$	Full	111	20.1	Give me a biography on Oscar Wilde, the scientist who discovered Quantum Mechanics, Schrödinger's Cat Thought Experiment.
ELI5$_{PIC}$	Full	146	12.5	Answer the following question(s): why it's common to have 87-octane gasoline in the US but it's almost always 95-octane in Europe?
AskHistorians$_{PIC}$	Full	158	19.2	In the original Star Wars: A New Hope, Obi-Wan Kenobi instructs R2-D2 to connect to the Imperial network to gain access to the whole system. Did the concept of an interconnected vast computer network exist in 1977?
ExpertQA$_{PIC}$	Full	152	13.5	Answer the question(s): What's the difference between modern and contemporary architecture?
FACTS$_{PIC}$	Partial	150	63.5	Explain the benefits of using mobile technology to improve healthcare management in both hi-income and low-income countries. {Context p}
XSUM$_{PIC}$	Partial	200	30.6	Summarize the following text in around 20-25 words. {Context p}

N = number of tasks; **C** = average number of input claims

Model Baselines

Open Models

Meta Llama 3.1 Instruct (8B, 70B)

Tulu 3 (8B, 70B)

Mistral AI Ministral 8B

Nous Research Hermes 3 (8B, 70B)

Reasoning Models

Qwen 3 (32B)

QwQ 3 (32B)

R1-Qwen (32B)

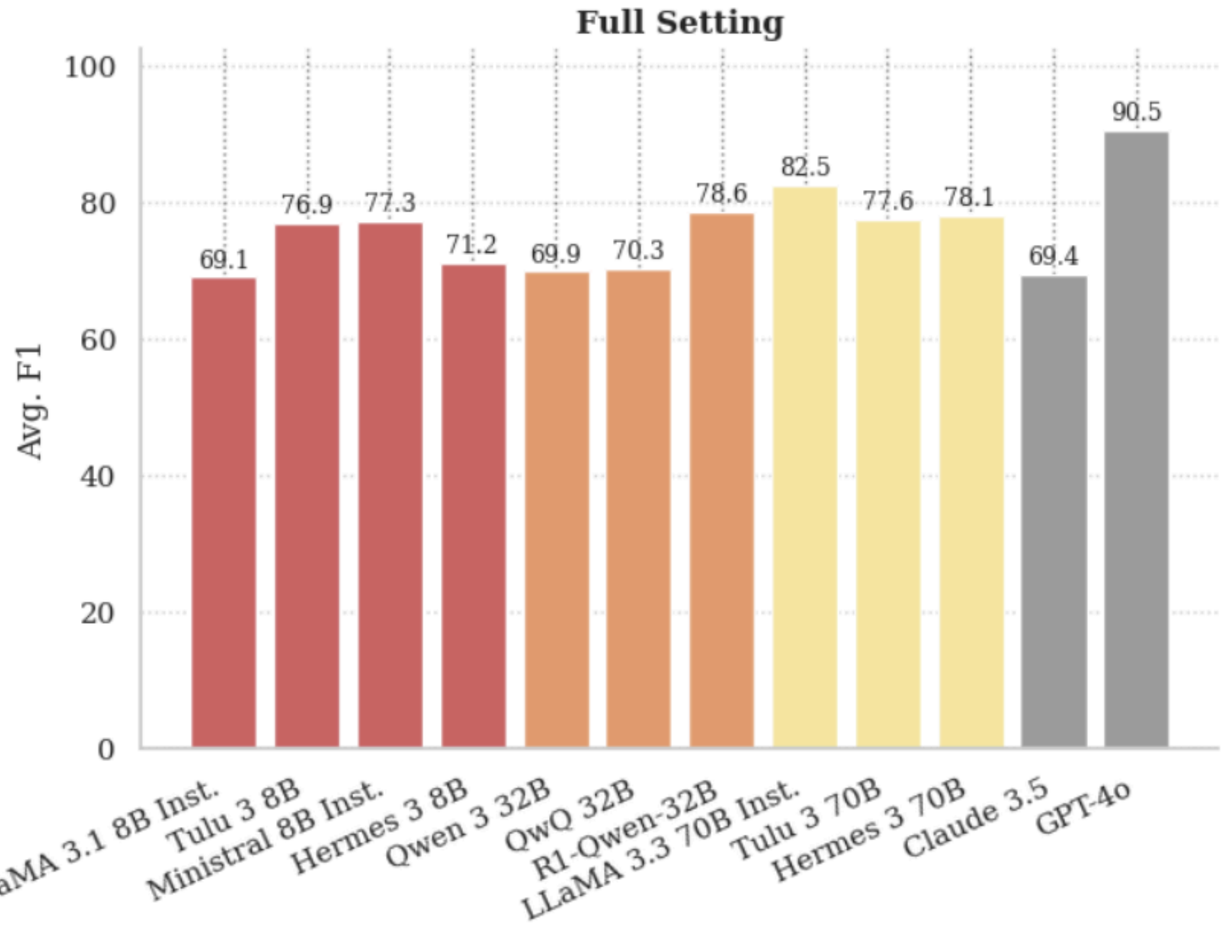
Closed Models

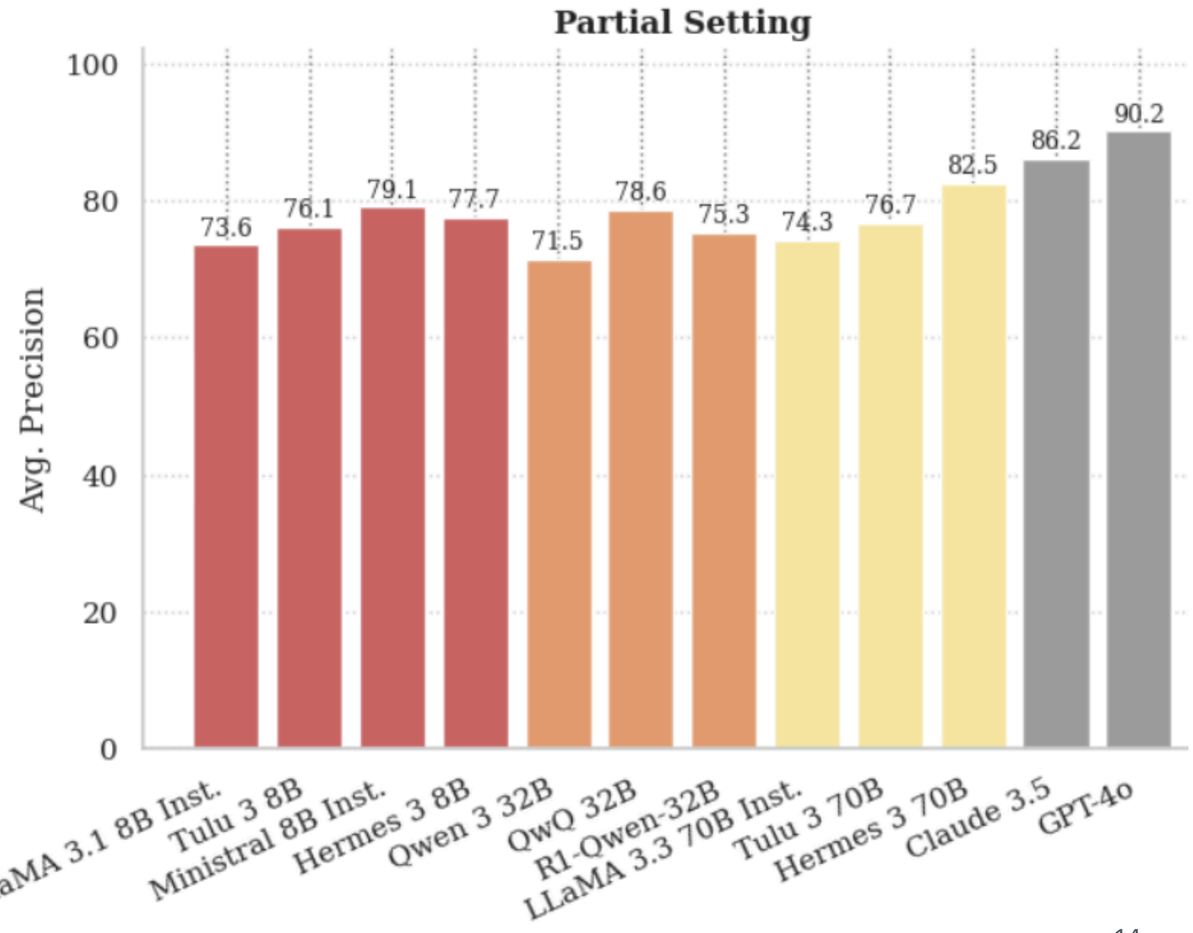
GPT-4

Claude 3.5 Sonnet

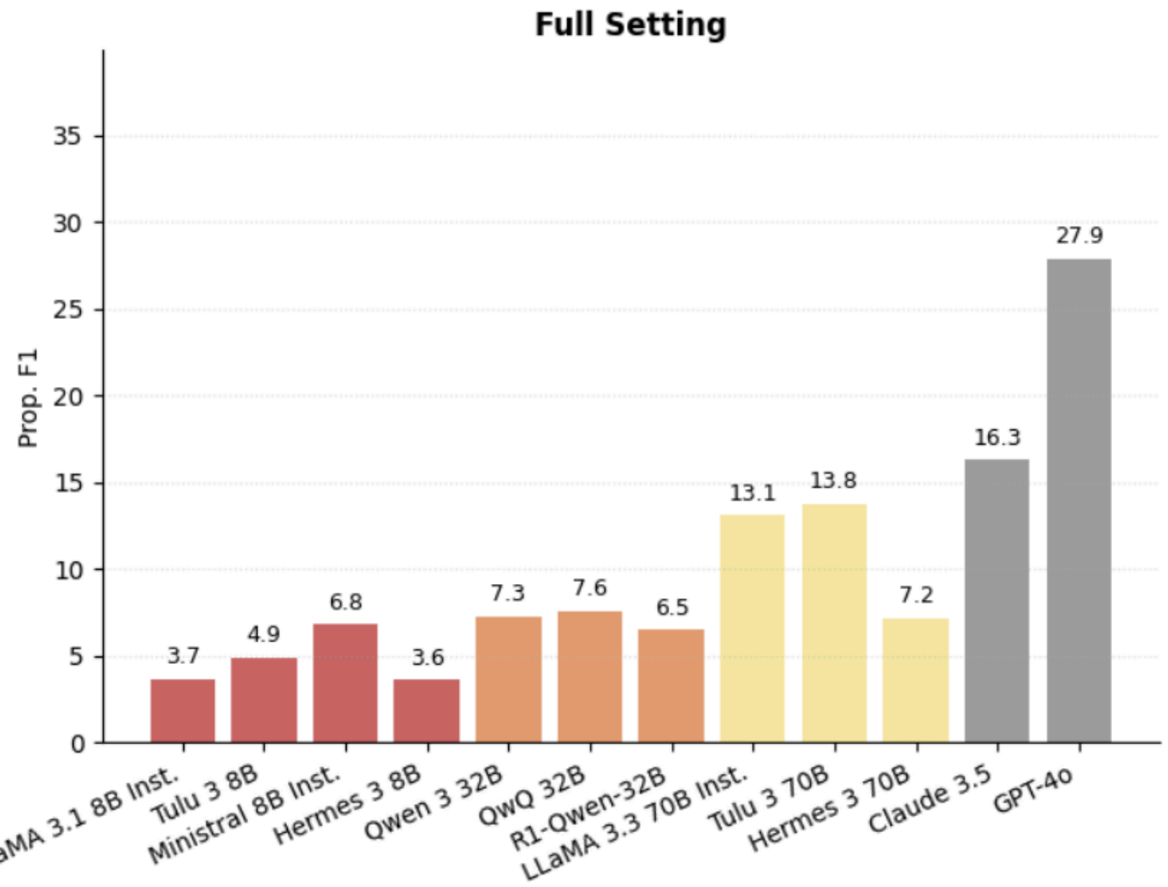
We evaluate many leading open and closed instruction-tuned models!

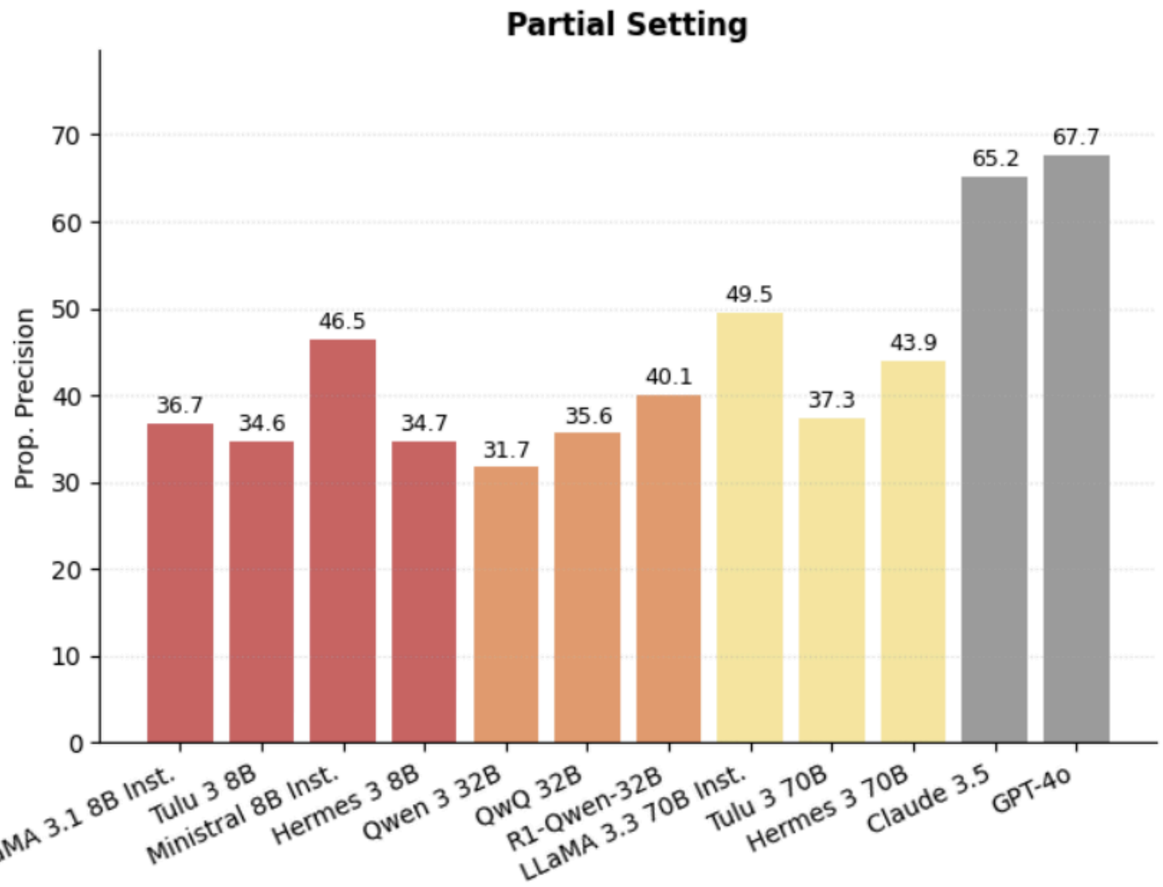
PIC-Bench Avg. Results





The ideal case is challenging…





Open-weight (8B) Open-weight (32B) Open-weight (70B) Proprietary

2. *PIC-LM*

How can we improve PIC ability in LMs?

PIC-LM Training Framework

We propose a training framework to produce PIC-LM 8B (initialized from Llama 3.1 8B Instruct)

It's not very complicated! Just SFT + normalized DPO on PIC-formatted data.

However, data quality and curation are very important at both stages!

Llama 3.1 8B Inst. PIC-LM 8B SFT PIC-LM 8B

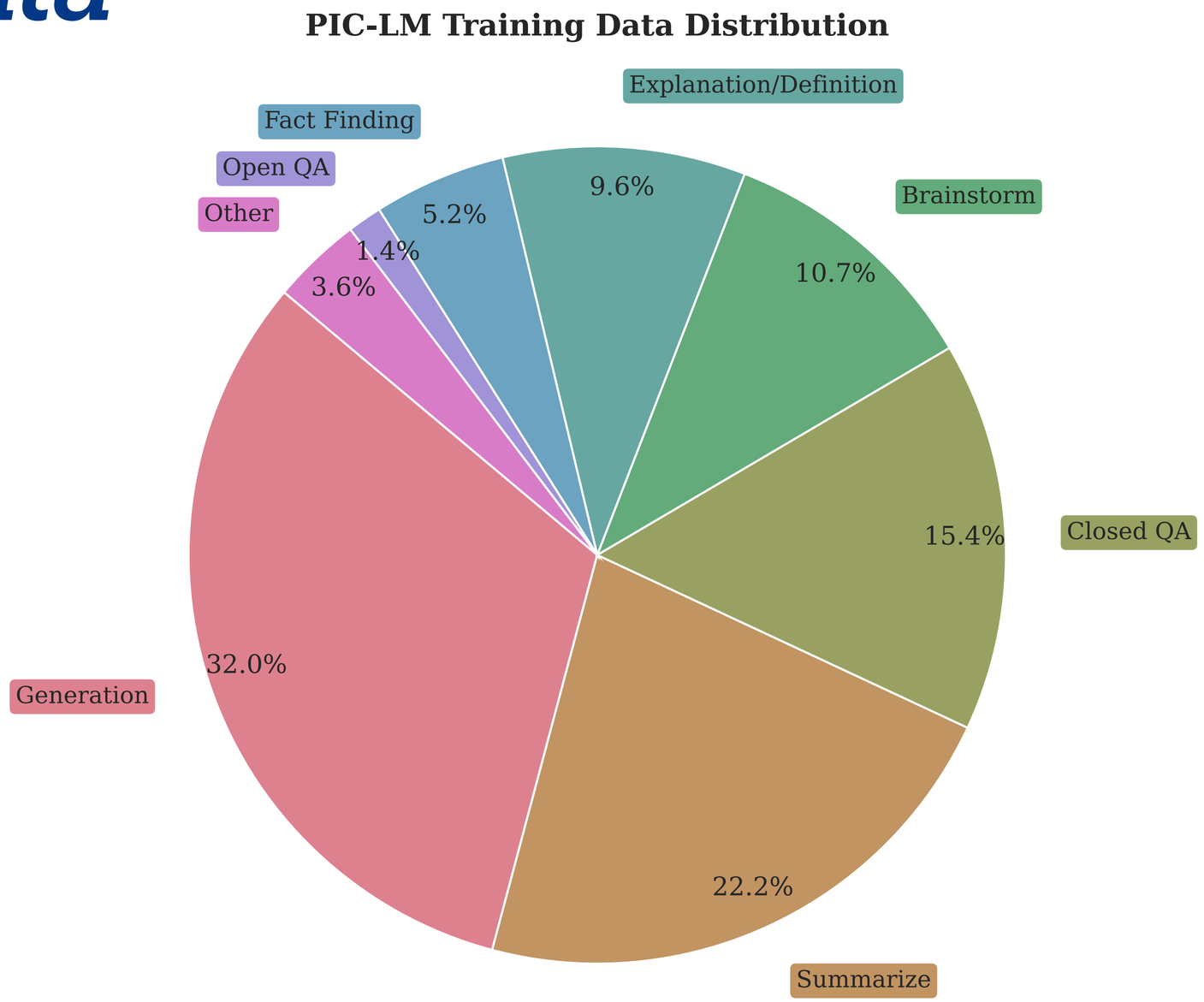
SFT Training Data

We choose high-quality, long-form instruction tuning data and reformat them to PIC.

General sources: No Robots, FLAN

Domain-specific: CNN news summarization, biography generation, FACTS Grounding long-form QA

Ablations show that both task diversity and PIC quality are important!

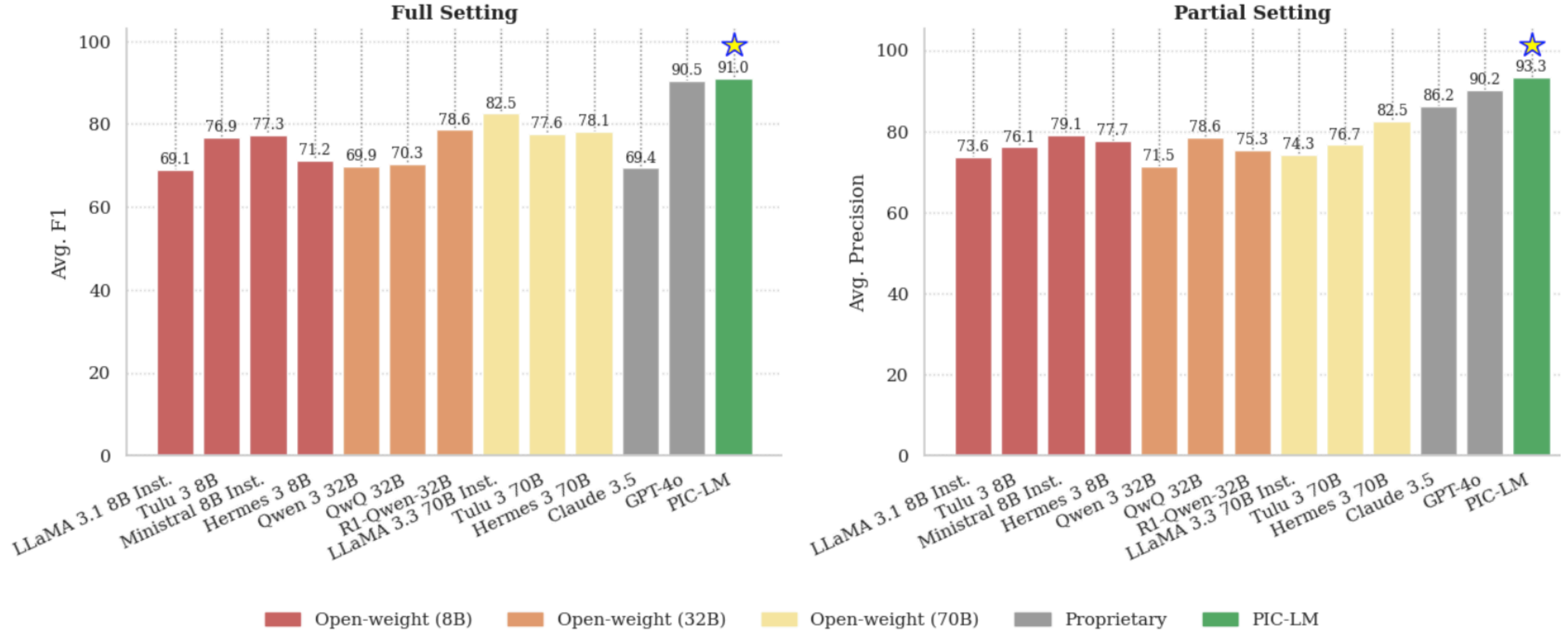


Preference Data Creation

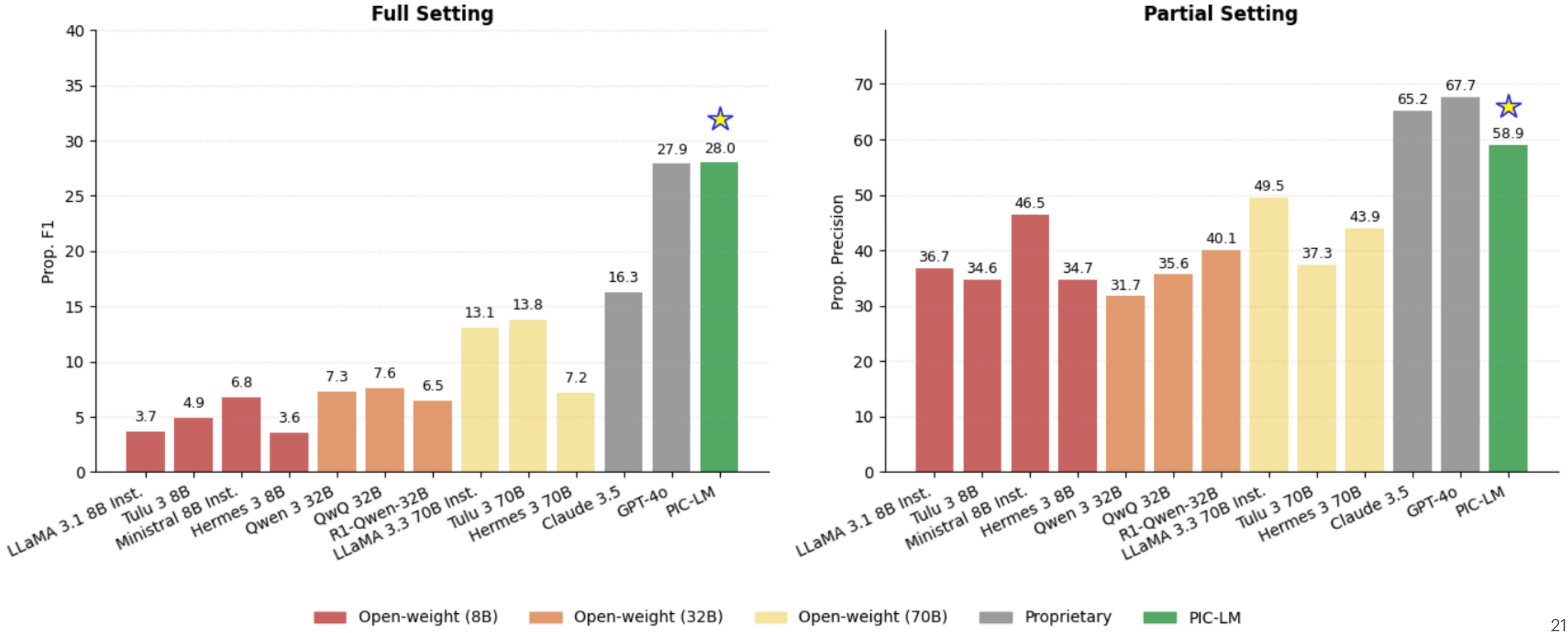
How to make PIC preference data construction **scalable**?

We introduce a **weakly supervised** preference data construction strategy that reuses PIC-LM SFT and our SFT data.

PIC-LM improves average PIC ability...



...Even in ideal-case settings!

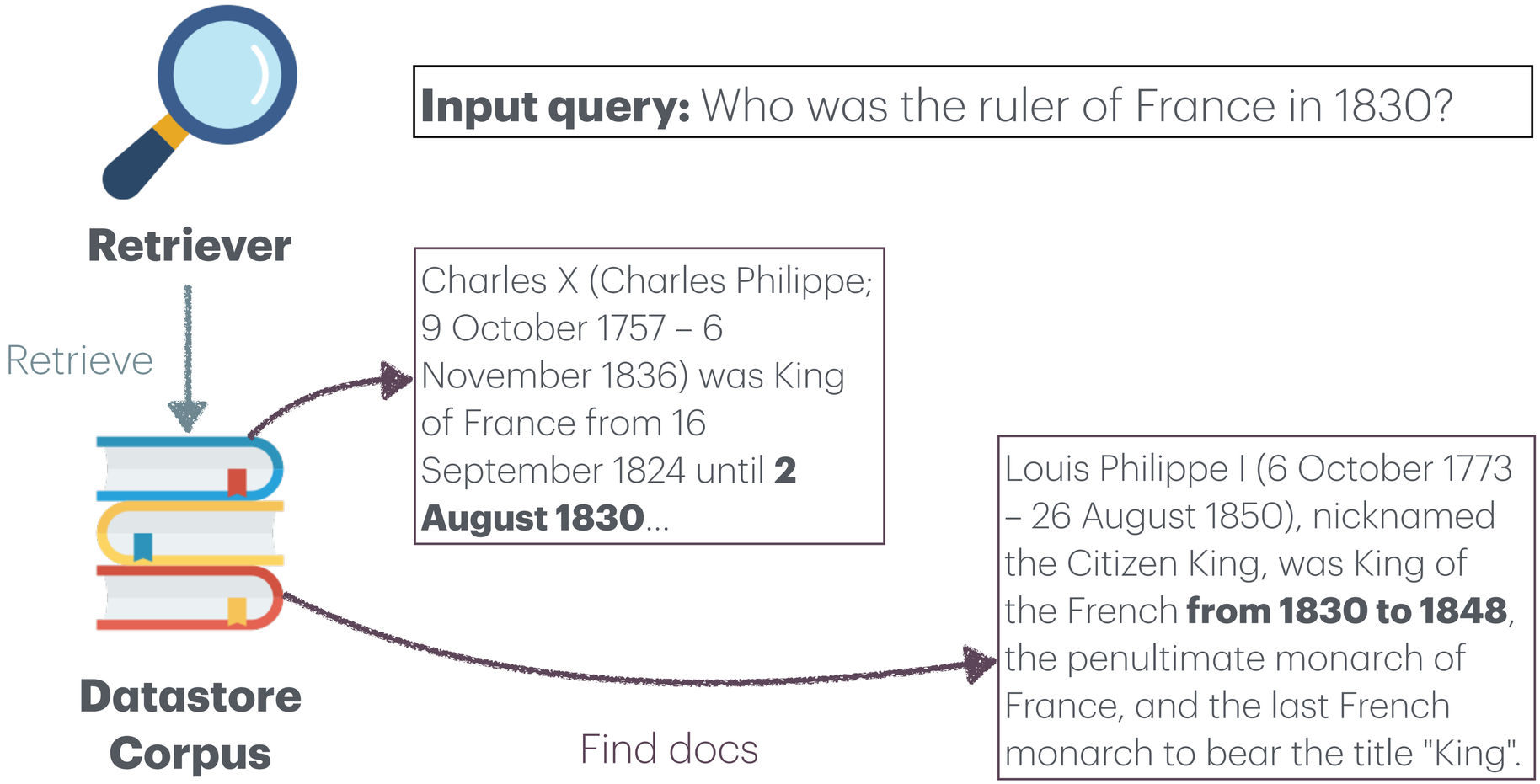


3. *PIC* Use Cases

Why should we care about better PIC ability?

Retrieval-Augmented Generation (RAG)

Goal: Supplement the input query with retrieved texts from a datastore (e.g., Wikipedia) to guide the LM's generation.



Retrieval-Augmented Generation (RAG)

We evaluate on **ASQA** [Stelmakh et al., 2022], an ambiguous long-form QA task, using retrieved Wikipedia passages from ALCE [Gao et al., 2025].

We pass in retrieved context at the granularity of decomposed **verifiable claims**, and report the **Exact Match (EM)**.

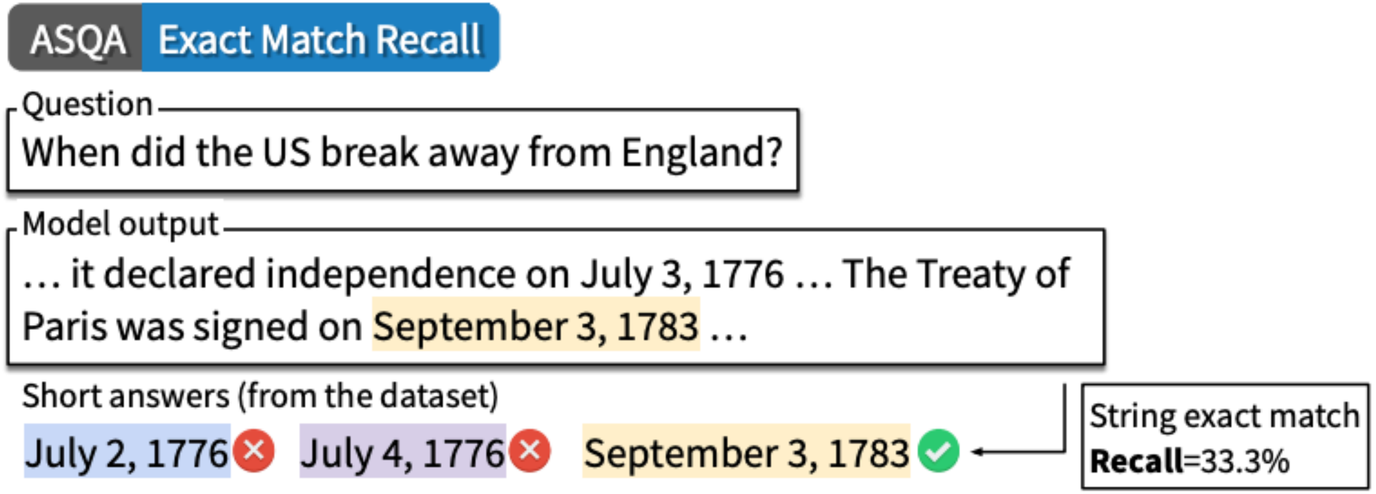
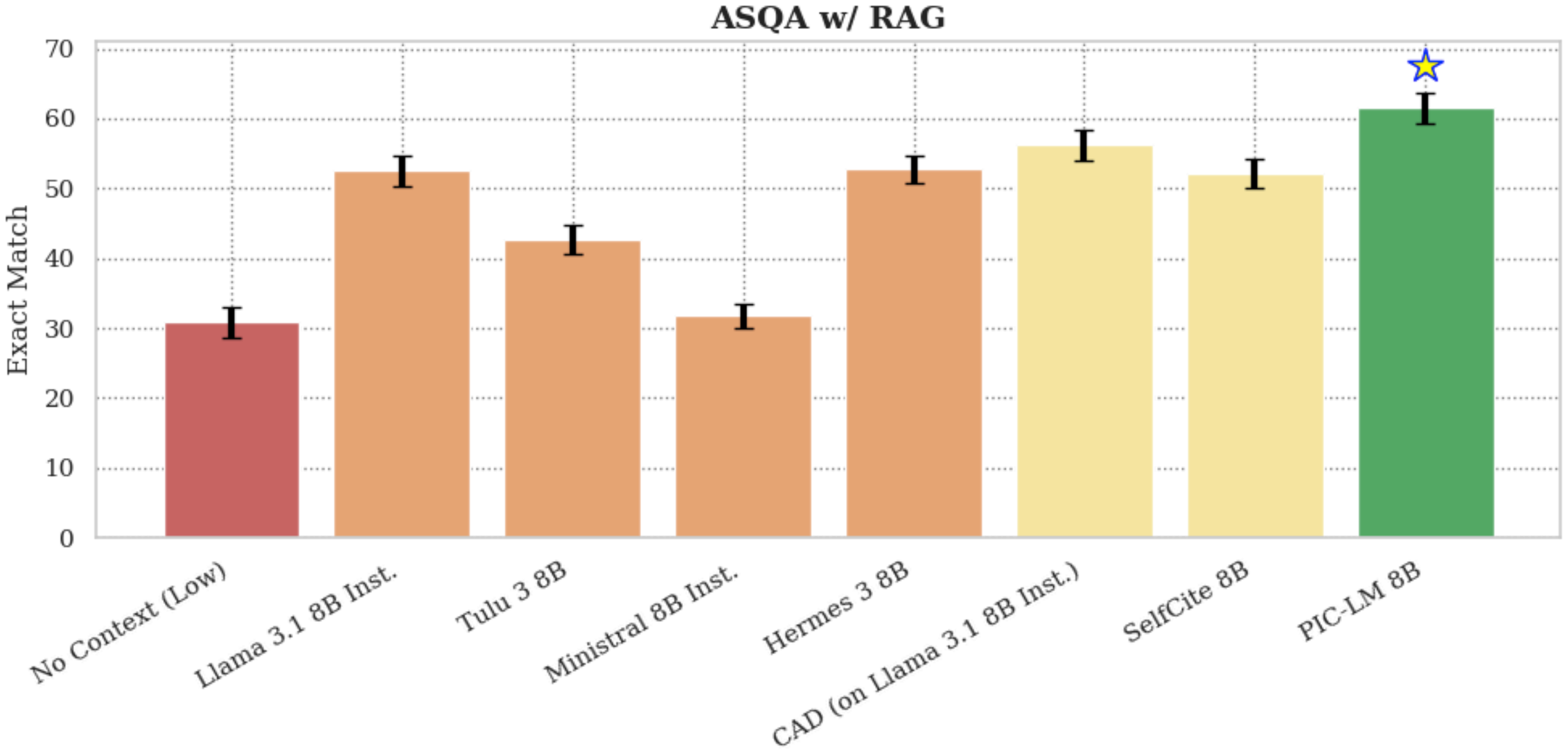


Figure from Gao et al., 2025.

Stelmakh et al., 2022. ASQA: Factoid questions meet long-form answers.

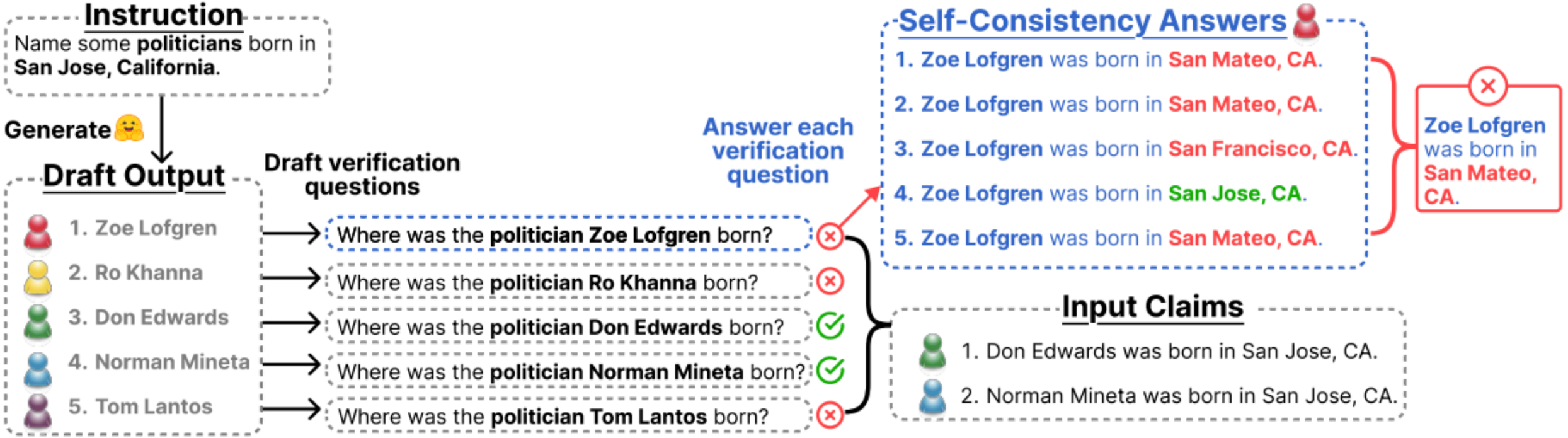
Gao et al., 2025. Enabling large language models to generate text from citations.

RAG Results


ASQA w/ RAG
Exact Match
70
60
50
40
30
20
10
0
No Context (Low)
Llama 3.1 8B Inst.
Tulu 3 8B
Ministral 8B Inst.
Hermes 3 8B
CAD (on Llama 3.1 8B Inst.)
SelfCite 8B
PIC-LM 8B

But RAG is still reliant on external claims...

Chain-of-Verification (CoVE)



Under CoVE, LMs can generate and verify its own high-quality claims!

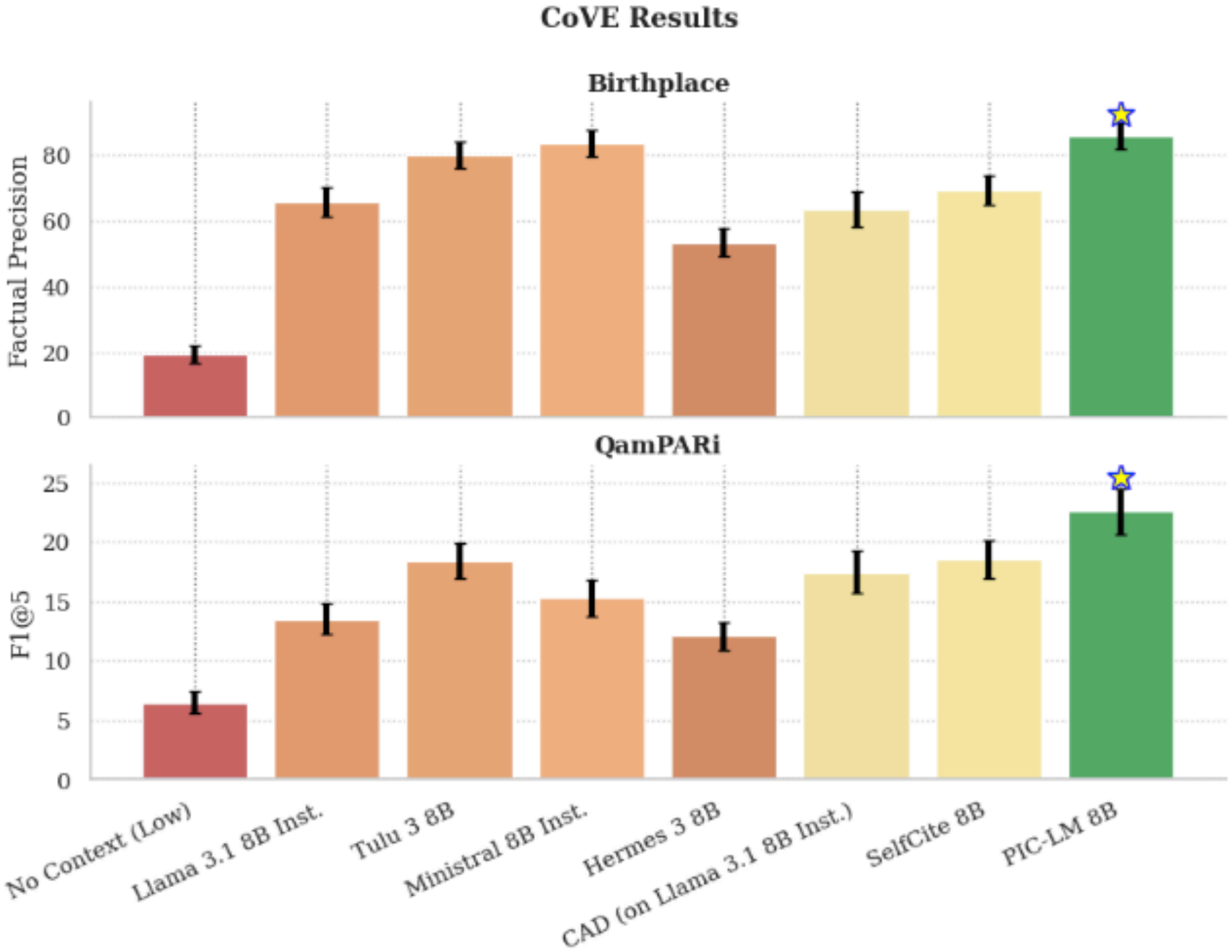
Chain-of-Verification (CoVE)

We consider two tasks:

- **Birthplace factoids** [Dhuliawala et al., 2023]: “Name some doctors born in San Jose, California.”
  - We report factual precision (checking against web search).
- **QAMPARI** [Amouyal et al., 2023]: “Which book had illustrations by Pauline Baynes?”.
  - We report F1@*k*, where *k*=5.

Dhuliawala et al., 2023. Chain-of-verification reduces hallucination in large language models.
Amouyal et al., 2023. QAMPARI: A Benchmark for open-domain questions with many answers.

CoVe Results



4. Key Takeaways

What's next?

Key Takeaways

- We frame PIC as a claim-level control problem for **faithfulness hallucination**.
- We present a new task formulation, a benchmark, initial post-training recipes, and several motivating use cases.
- Despite its simple premise, PIC is still an open problem!

Thank you!!

☺

ArXiv