COMPUTER SCIENCE
& ENGINEERING
UNIVERSITY OF MICHIGAN

ICLR

LitCab: Lightweight Language Model Calibration over Short- and Long-form Responses

Xin Liu, Muhammad Khalifa, Lu Wang
University of Michigan

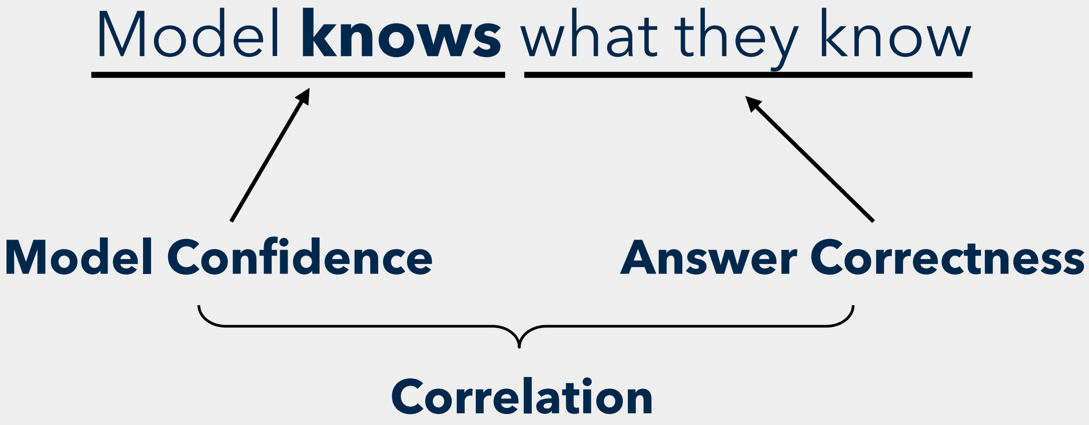
Background

What is model calibration…

Model **knows** what they know

Background

What is model calibration…



Model **knows** what they know

Model Confidence

Answer Correctness

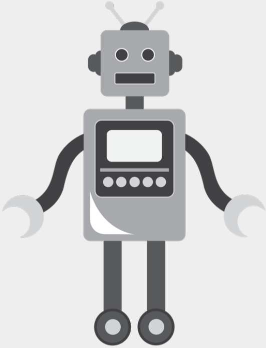
Correlation

Background

What is model calibration...

Model knows what they know

Who is the president of the United States in 2023?



Donald Trump (Confidence 0.95)

Joe Biden (Confidence 0.05)

Donald Trump (Confidence 0.05)

Joe Biden (Confidence 0.95)

Background

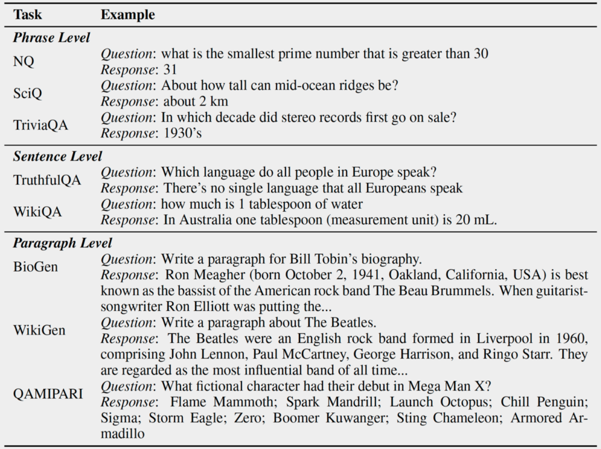
Why is model calibration so important?

It makes the model trustworthy!

- The well-calibrated confidence enables users to make **informed decisions** about whether to trust its outputs.

- A well-calibrated model can **refuse to answer** questions if it is not certain.

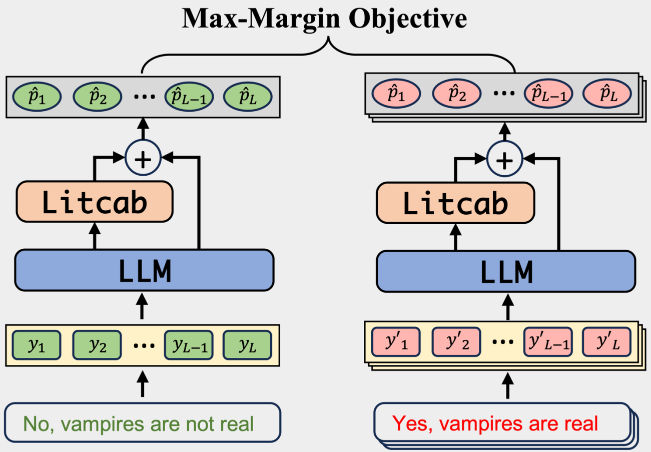
Roadmap



Task	Example
Phrase Level	
NQ	*Question*: what is the smallest prime number that is greater than 30 *Response*: 31
SciQ	*Question*: About how tall can mid-ocean ridges be? *Response*: about 2 km
TriviaQA	*Question*: In which decade did stereo records first go on sale? *Response*: 1930's
Sentence Level	
TruthfulQA	*Question*: Which language do all people in Europe speak? *Response*: There's no single language that all Europeans speak
WikiQA	*Question*: how much is 1 tablespoon of water *Response*: In Australia one tablespoon (measurement unit) is 20 mL.
Paragraph Level	
BioGen	*Question*: Write a paragraph for Bill Tobin's biography. *Response*: Ron Meagher (born October 2, 1941, Oakland, California, USA) is best known as the bassist of the American rock band The Beau Brummels. When guitarist-songwriter Ron Elliott was putting the...
WikiGen	*Question*: Write a paragraph about The Beatles. *Response*: The Beatles were an English rock band formed in Liverpool in 1960, comprising John Lennon, Paul McCartney, George Harrison, and Ringo Starr. They are regarded as the most influential band of all time...
QAMIPARI	*Question*: What fictional character had their debut in Mega Man X? *Response*: Flame Mammoth; Spark Mandrill; Launch Octopus; Chill Penguin; Sigma; Storm Eagle; Zero; Boomer Kuwanger; Sting Chameleon; Armored Armadillo

CaT Benchmark

A **Ca**libration Evalua**T**ion Benchmark, in particular **long-form generation**



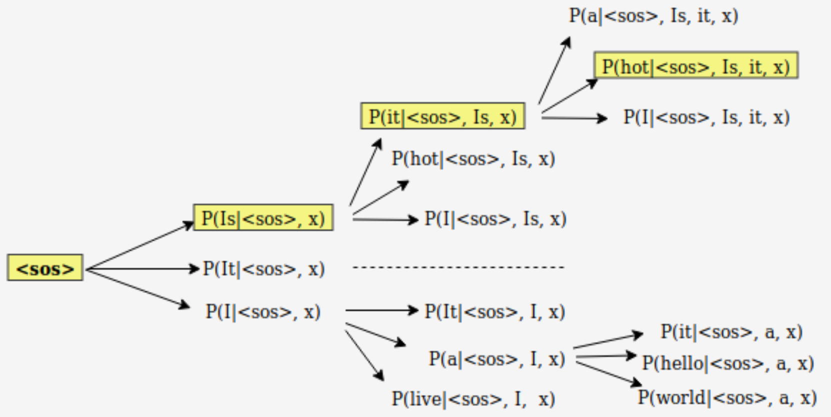
Max-Margin Objective
$\hat{p}_1$ $\hat{p}_2$ ⋯ $\hat{p}_{L-1}$ $\hat{p}_L$
+
Litcab
LLM
y_1 y_2 ⋯ y_{L-1} y_L
No, vampires are not real
$\hat{p}_1$ $\hat{p}_2$ ⋯ $\hat{p}_{L-1}$ $\hat{p}_L$
+
Litcab
LLM
y'_1 y'_2 ⋯ y'_{L-1} y'_L
Yes, vampires are real

LitCab

Lightweigh**t** **Cal**i**b**ration of Language Models

CaT Benchmark

Confidence Estimation



P(a|<sos>, Is, it, x)

P(hot|<sos>, Is, it, x)

P(it|<sos>, Is, x)

P(I|<sos>, Is, it, x)

P(hot|<sos>, Is, x)

P(Is|<sos>, x)

P(I|<sos>, Is, x)

<sos>

P(It|<sos>, x)

P(I|<sos>, x)

P(It|<sos>, I, x)

P(it|<sos>, a, x)

P(a|<sos>, I, x)

P(hello|<sos>, a, x)

P(live|<sos>, I, x)

P(world|<sos>, a, x)

Retrieved from https://towardsdatascience.com/word-sequence-decoding-in-seq2seq-architectures-d102000344ad

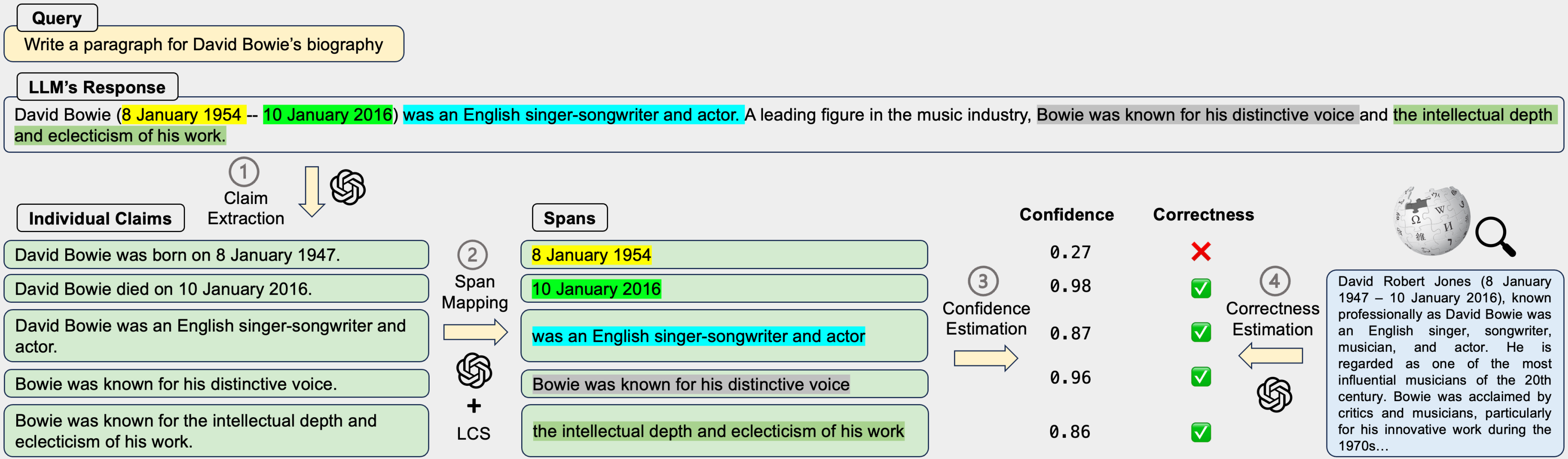
Model’s Confidence:

$$p(y|x) = \sqrt[L]{\prod_{t=1}^{L} p(y_t|x, y_{<t})}$$

This works for short model generations, i.e. **phrases** and **sentences.**
*How about **long-form generations**?*

CaT Benchmark

Claim-level Calibration Evaluation for Long-form Generations



Query

Write a paragraph for David Bowie’s biography

LLM’s Response

David Bowie (8 January 1954 -- 10 January 2016) was an English singer-songwriter and actor. A leading figure in the music industry, Bowie was known for his distinctive voice and the intellectual depth and eclecticism of his work.

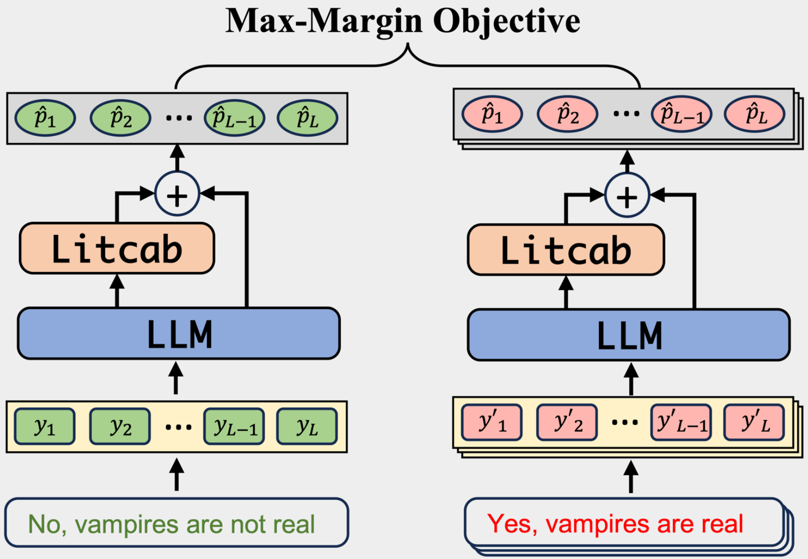
Individual Claims	Spans	Confidence	Correctness
David Bowie was born on 8 January 1947.	8 January 1954	0.27	✗
David Bowie died on 10 January 2016.	10 January 2016	0.98	✓
David Bowie was an English singer-songwriter and actor.	was an English singer-songwriter and actor	0.87	✓
Bowie was known for his distinctive voice.	Bowie was known for his distinctive voice	0.96	✓
Bowie was known for the intellectual depth and eclecticism of his work.	the intellectual depth and eclecticism of his work	0.86	✓

David Robert Jones (8 January 1947 – 10 January 2016), known professionally as David Bowie was an English singer, songwriter, musician, and actor. He is regarded as one of the most influential musicians of the 20th century. Bowie was acclaimed by critics and musicians, particularly for his innovative work during the 1970s…

Estimating the confidence and correctness of claims.

LitCab

- **Architecture:** a single linear layer processes the model's top-layer hidden states as input and predicts a bias term for the model-generated logits.
- **Training Objective:** Max-Margin Objective, which increases the likelihood of positive samples, while decreases that of negative samples.


Max-Margin Objective
$\hat{p}_1$ $\hat{p}_2$ … $\hat{p}_{L-1}$ $\hat{p}_L$
+
Litcab
LLM
y_1 y_2 … y_{L-1} y_L
No, vampires are not real
$\hat{p}_1$ $\hat{p}_2$ … $\hat{p}_{L-1}$ $\hat{p}_L$
+
Litcab
LLM
y'_1 y'_2 … y'_{L-1} y'_L
Yes, vampires are real

$$L(x,y) = \sum_{y' \in I(y)} \max(0, 1 + \hat{p}(\underline{y'}|x) - \hat{p}(\underline{y}|x))$$

Negative Samples **Positive Samples**

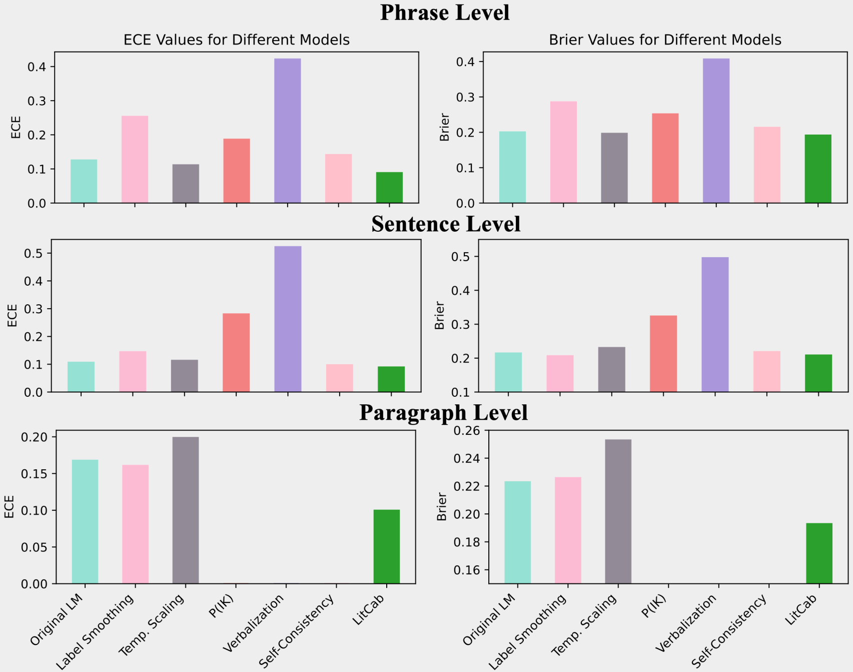
Experiment Settings

- **Benchmark: CaT**
  - **Phrase-level tasks:** NQ, SciQ, TriviaQA
  - **Sentence-level tasks:** TruthfulQA, WikiQA
  - **Paragraph-level tasks:** BioGen, WikiGen
- **Calibration Metrics: ECE** (Expected Calibration Error), **Brier** Score, both of which are better when lower.
- **Basic LLM: Llama2 7b**

Results of LitCab

#1 Takeaway: Compared with methods that requires tuning additional parameters, LitCab shows better calibration performance.

#2 Takeaway: LitCab performs as well as or better than the strongest comparison self-consistency, while achieving **superior inference efficiency.**



Phrase Level

Sentence Level

Paragraph Level

Thank you!