Fact Recall, Heuristics or Pure Guesswork?

Precise Interpretations of Language Models for Fact Completion

Denitsa Saynova*, Lovisa Hagström*, Moa Johansson, Richard Johansson, Marco Kuhlmann

Fact Completion

What is fact completion

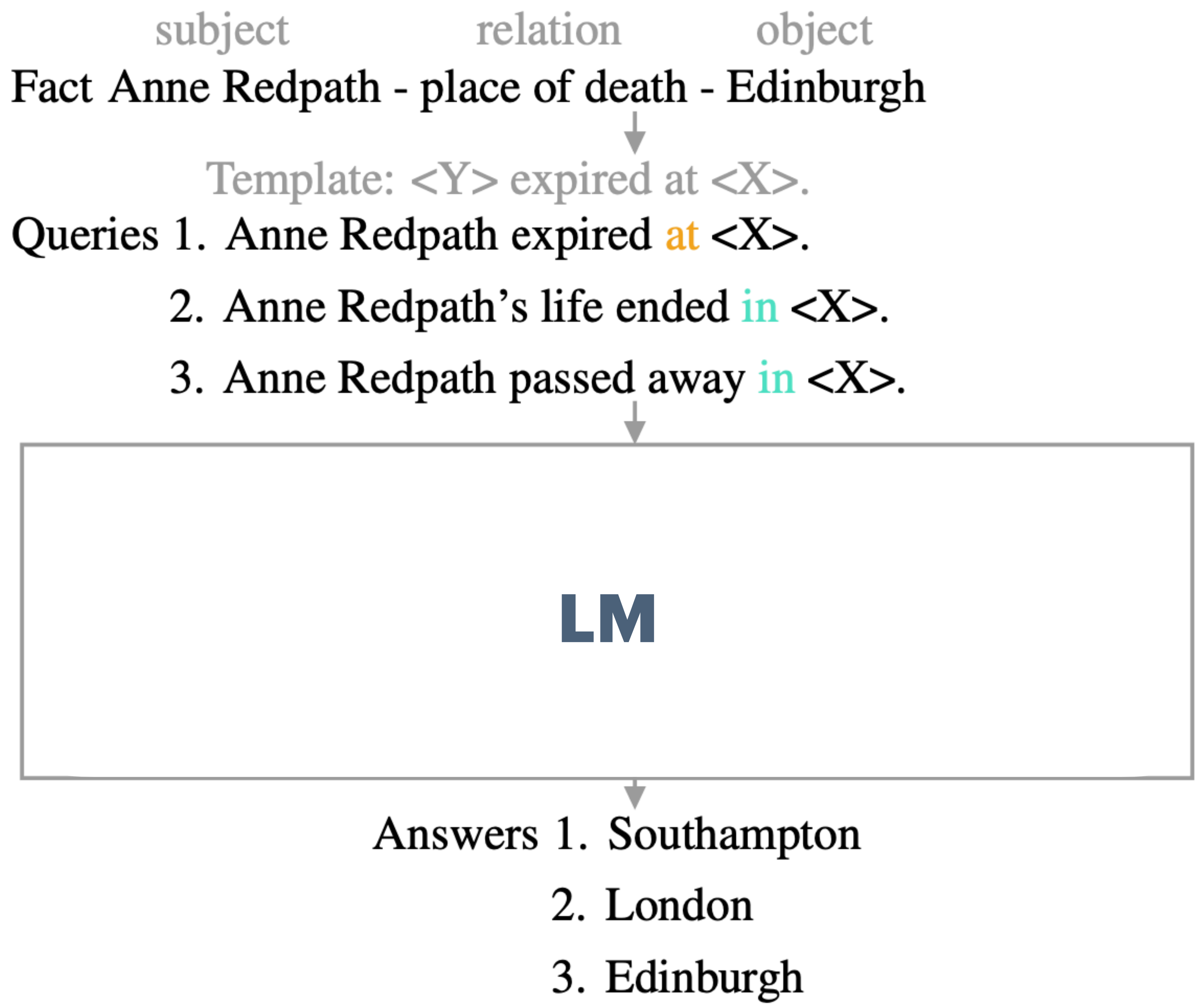
Subject	Relation	Object
Tokyo	**capital of**	**Japan**

Tokyo is the capital of → LM → Japan

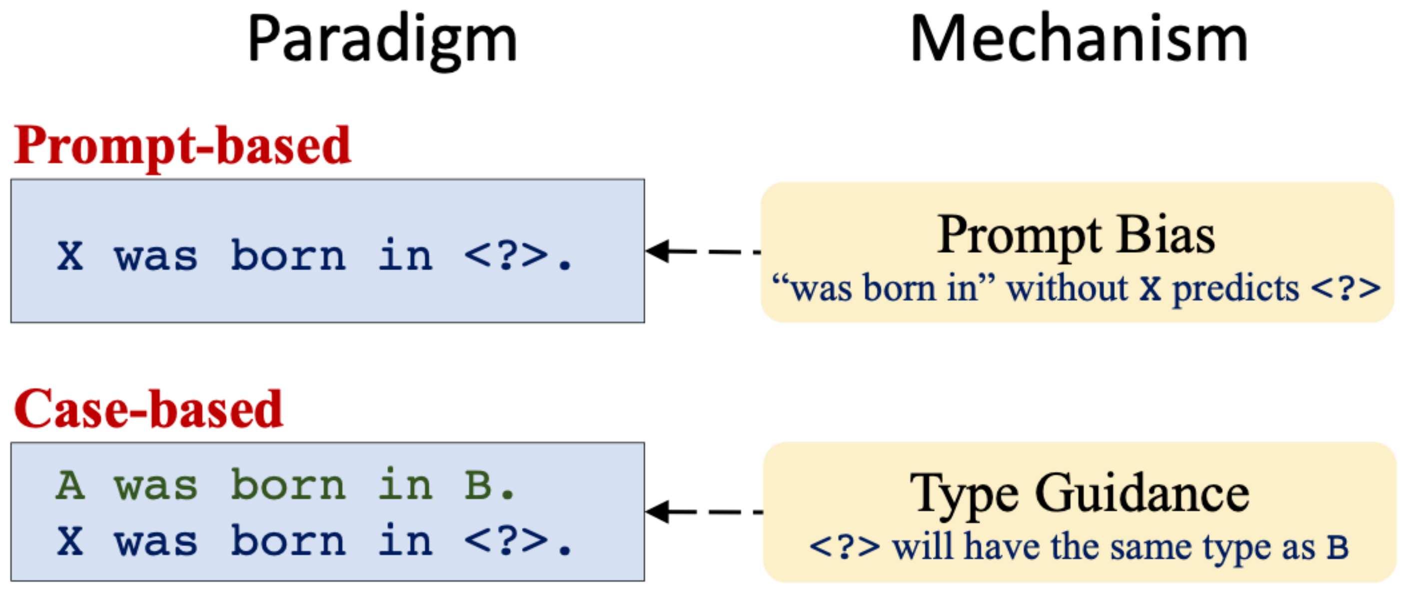
What we know about how fact completion is performed

- Mechanistic interpretability with interchange interventions
  - **Goal:** to identify model states that are involved in fact-related inference
  - **Assumption:** an accurate prediction indicates a memorised fact
  - **Main finding:** Last subject token, mid-layer MLP states act as a key-value store for facts

What we know about LM behaviour for fact prediction



The Effect of Scaling, Retrieval Augmentation and Form on the Factual Consistency of Language Models (Hagström et al., EMNLP 2023)



Knowledgeable or Educated Guess? Revisiting Language Models as Knowledge Bases (Cao et al., ACL-IJCNLP 2021)

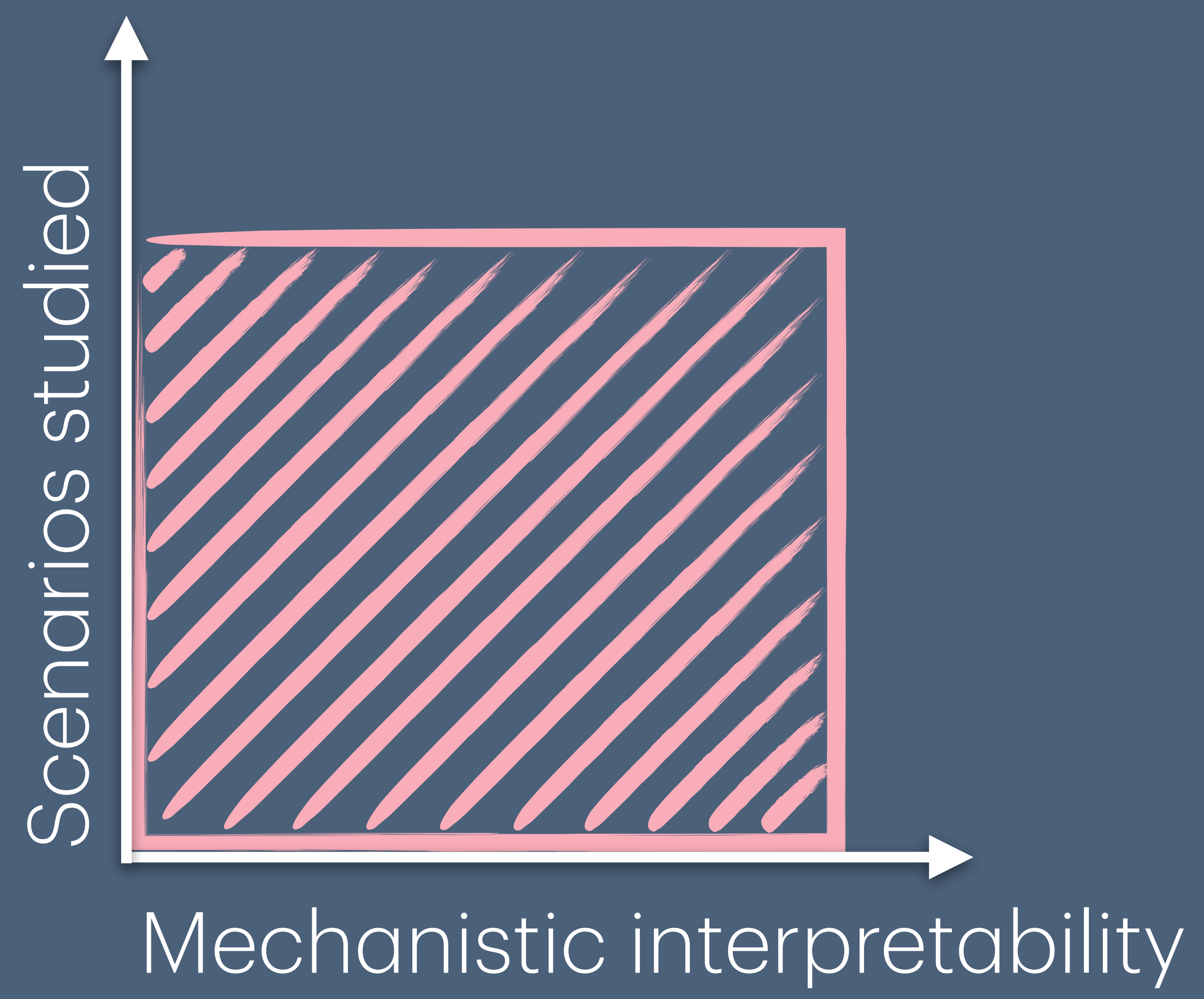
Article: Jung Lee is a well-known **French** writer who was **born in Paris**. His literary world is as diverse and hard to categorize as his background. He has lived in both urban and rural areas, deep in the mountains and in the seaside towns and has developed a wide range of interests from the tradition of Confucian culture to advertising.

Generated Summary: Jung Lee is one of **South Korea's** best-known writers.

When Do Pre-Training Biases Propagate to Downstream Tasks? A Case Study in Text Summarization (Ladhak et al., EACL 2023)

Our proposed framework

Precise Identification of Scenarios for Model behavior (PrISM)



Exact fact recall

Tokyo is the capital city of **Japan**

Heuristics recall

Kye Ji-Su, a citizen of **South Korea**

Both of these predictions are accurate.

Does that mean that they are equal in terms of fact recall?

Previous work
Yes.

Our work
No.

PrISM datasets for precise studies of prediction scenarios

- Our datasets are **model-specific** and aim to **separate different prediction scenarios**.
- An inspection of the 1,209 samples from CounterFact (a diagnostic dataset frequently used for model interpretations) reveals **~900 potentially problematic samples**.

Heuristics recall

Giuseppe Angeli, who has a citizenship of **Italy**

MacApp, a product created by **Apple**

[X] professionally plays the sport of ice **hockey**

Building a PrISM dataset

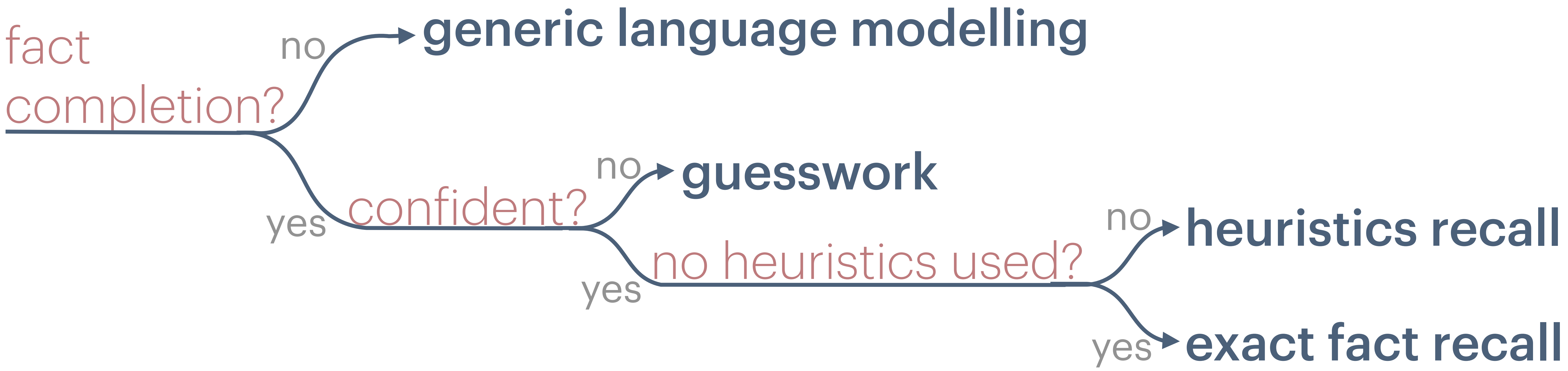
generic language modelling

guesswork

heuristics recall

exact fact recall

Building a PrISM dataset



Building a PrISM dataset

fact completion?

no → **generic language modelling**

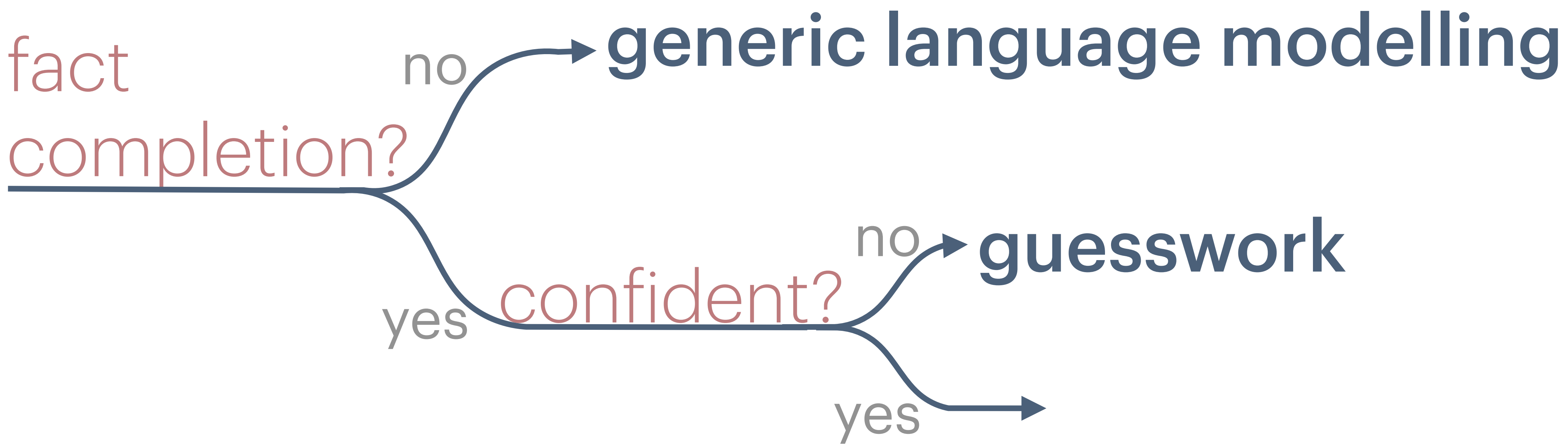
Kun-Woo Paik is also a regular guest artist at **the**

yes →

Eksi Ekso originated in **Russia**

Does the prompt and the model's prediction correspond to the setting of a model completing a fact?

Building a PrISM dataset

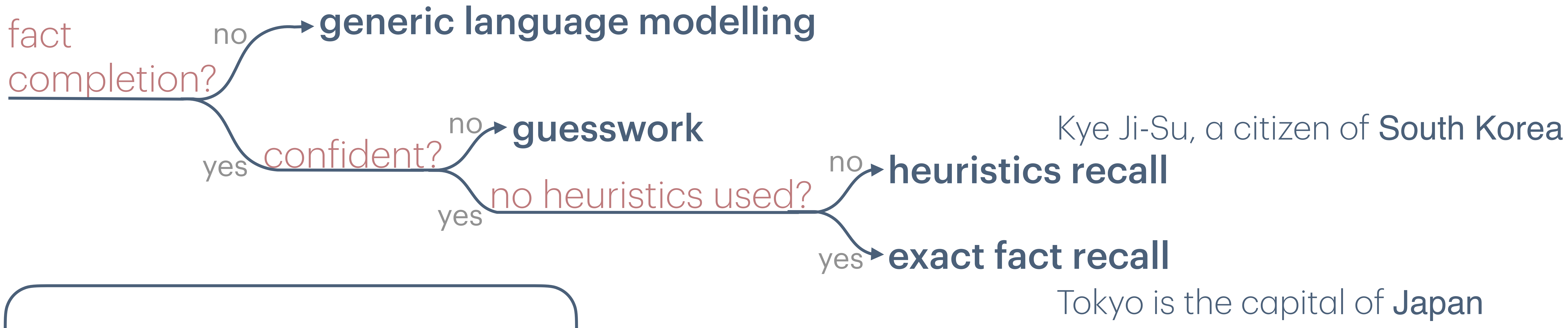


Eksi Ekso, a citizen of **Sweden** 🤷‍♀️
Eksi Ekso originated in **Russia**

Kye Ji-Su, a citizen of **South Korea**

Is the prediction confident? We proxy model confidence by consistency in the face of semantically equivalent queries.

Building a PrISM dataset



Is the prediction based on the exact factual information expressed in the prompt rather than heuristics?

Person name bias Kye Ji-Su is a common name in **South Korea**

Prompt bias [X] was produced by **Apple**

Lexical overlap Nokia cellphone was created by **Nokia**

Building a PrISM dataset

sample =
(query, prediction)

Generic language modelling

Sample sentences from Wikipedia starting with a subject. Discard sentences for which the continuation begins with a capital letter or number.

Guesswork

Populate fact prompt templates with subjects and objects from WikiData. Retain samples for which the prediction is a valid object but unconfident.

Heuristics recall

Populate fact prompt templates with synthetic fact tuples from a name generator. Samples corresponding to confident predictions are retained.

Exact fact recall

Populate fact prompt templates with subjects and objects from WikiData. Samples that are 1) confident, 2) not corresponding to any bias, 3) corresponding to a fact likely memorized by the LM, and 4) correct.

PrISM dataset

GPT-2 XL samples

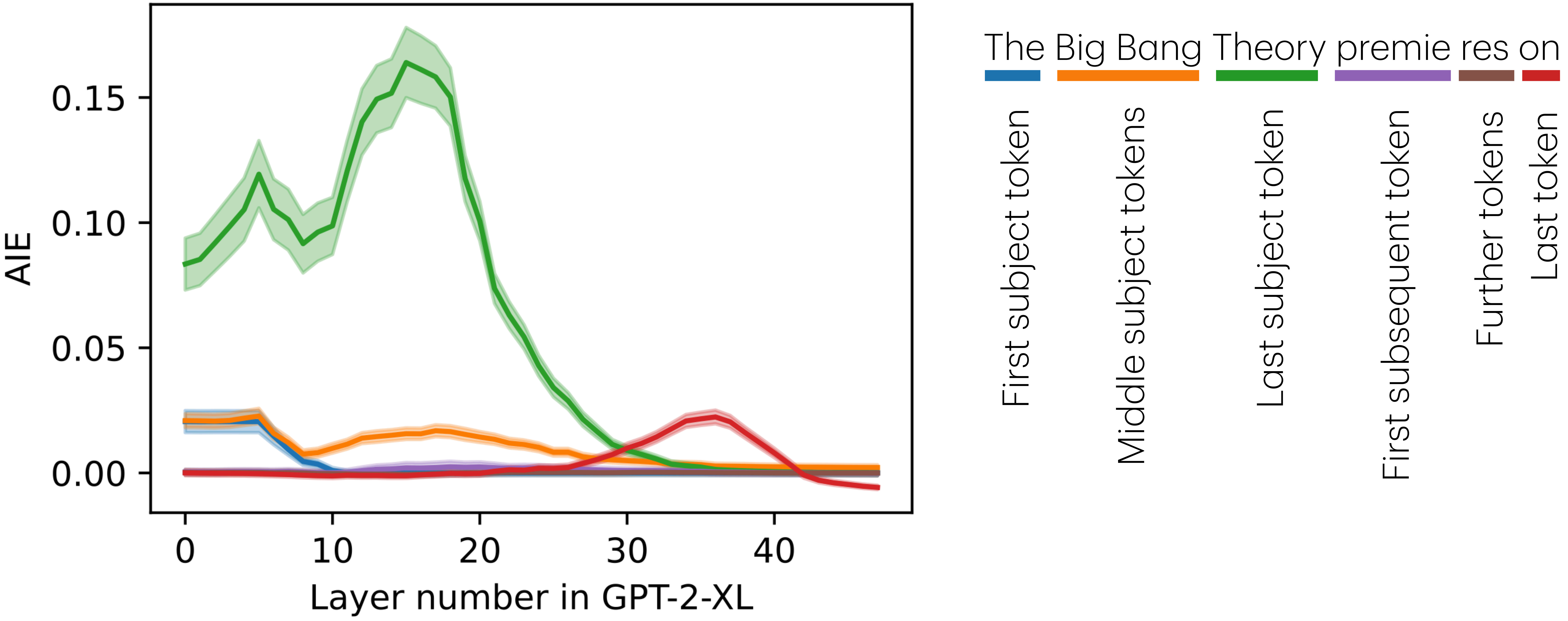
Scenario	Prompt	Prediction	Gold label	Conf	Pop	Bias
generic LM	Nara also enjoyed success in	the	singles	-	-	-
generic LM	Benjamin later joined a number of	other	clubs	-	-	-
guesswork	Sonar Kollektiv originated in	Russia	Berlin	1	215	-
guesswork	Joseph Clay was originally from	Ohio	Philadelphia	1	273	-
heuristics	Serok Nuvrome, a citizen of	Ukraine	-	6	0	name
heuristics	Balo Windhair has a citizenship of	Canada	-	5	0	prompt
exact fact	Thomas Ong is a citizen of	Singapore	Singapore	7	1418	none
exact fact	Shibuya-kei, that was created in	Japan	Japan	8	5933	none

PrISM datasets

Scenario	GPT-2 XL #samples (#fact tuples)	Llama 2 7B #samples (#fact tuples)	Llama 2 13B #samples (#fact tuples)
Generic LM	1000 (-)	1000 (-)	1000 (-)
Guesswork	3282 (3181)	2917 (2846)	2822 (2220)↓
Heuristics	8352 (1868)	8414 (1960)	9224 (2062)↑
Exact fact	1322 (191)	5481 (580)	5995 (601)↑

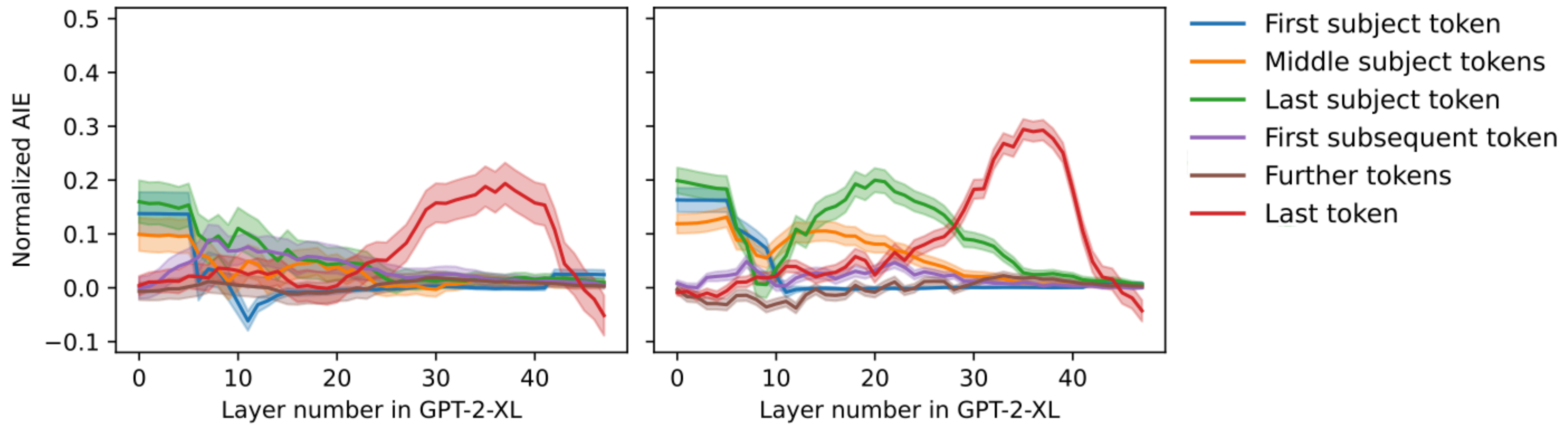
Causal Tracing Results

Causal Tracing Method



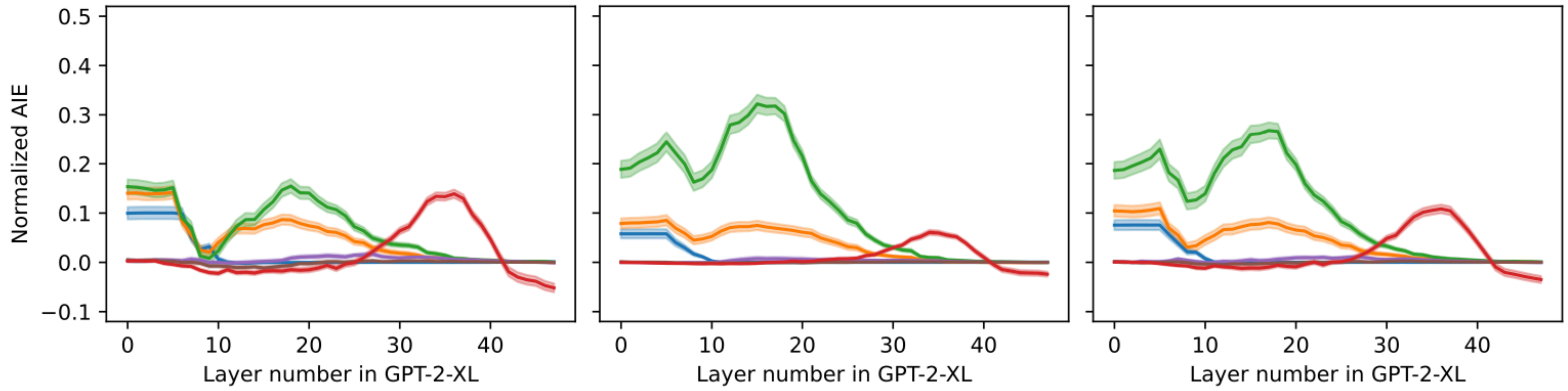
Normalisation

CT and PrISM



(a) Generic language modeling samples.

(b) Guesswork samples.



(c) Heuristics recall samples.

(d) Exact fact recall samples.

(e) Combined samples.

Summary

Accuracy does not indicate a consistent recall pattern

PrISM: Provides a taxonomy of four prediction scenarios

- Exact fact recall
- Heuristics recall
- Guesswork
- Generic language modelling

CT indicates different recall mechanisms for these scenarios

Fact Recall, Heuristics or Pure Guesswork?

Precise Interpretations of Language Models for Fact Completion

Preprint: https://arxiv.org/abs/2410.14405

Denitsa Saynova*, Lovisa Hagström*,
Moa Johansson, Richard Johansson, Marco Kuhlmann

NLP@DSAI

https://dsai-nlp.github.io/

CHALMERS

UNIVERSITY OF GOTHENBURG

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

WASP—HS