Heidelberg University
Cross-Temporal NLP

Knowledge Conflicts for LLMs

A Survey by

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, Wei Xu

Bohdana Ivakhnenko

November 11, 2025

Some slides are inspired by the presentation of Rongwu Xu and Zehan Qi

Knowledge Conflicts for LLMs

1. Introduction
2. Model Behaviours
3. Inter-Context Conflict
4. Intra-Memory Conflict
5. Context-Memory Conflict
6. Overview of Solutions
7. Considerations
8. Critique
9. Discussion Questions

Introduction

Overview

- Knowledge
  - Humans
  - LLMs
- Knowledge Conflicts
  - In humans
  - In knowledge graphs
  - In models

Knowledge

The Analysis of Knowledge

Is knowledge all there is?, blog

Knowledge

Philosophy

The Tripartite Analysis of Knowledge:
Subject S knows a proposition p iff

- p is true;
- S believes that p;
- S is justified in believing that p.

The Analysis of Knowledge

Is knowledge all there is?, blog

Knowledge

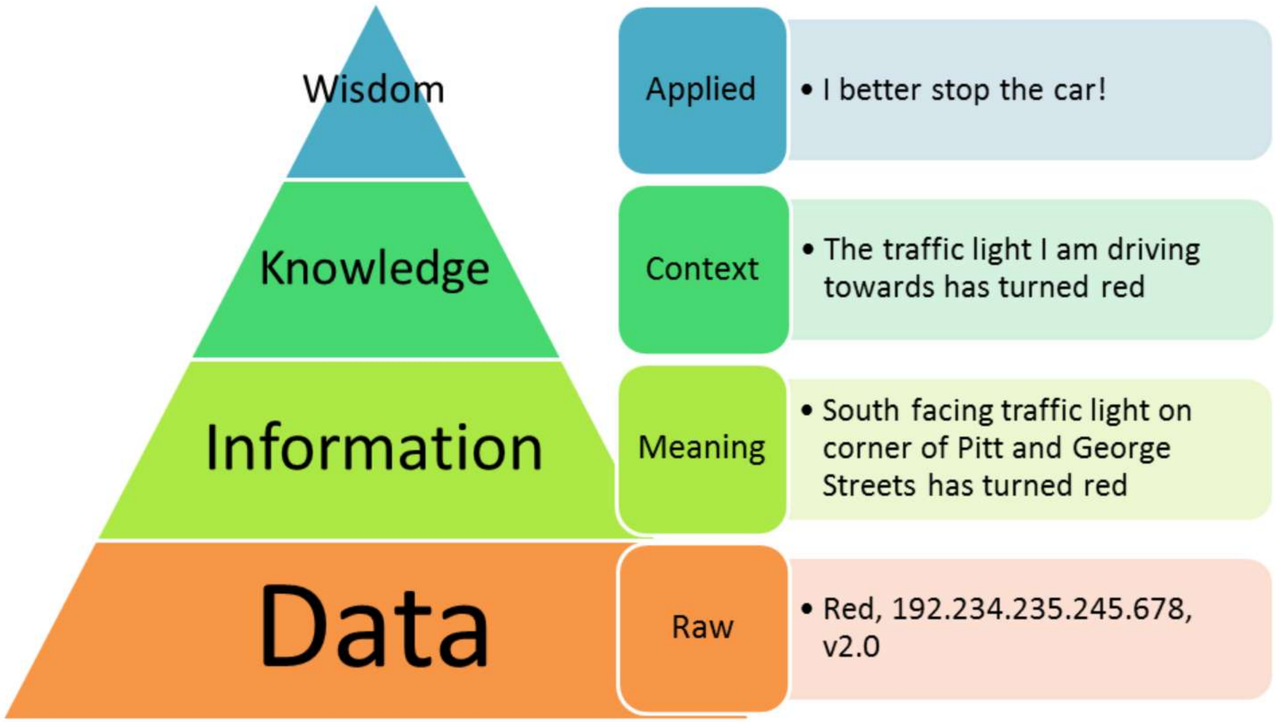
Philosophy

The Tripartite Analysis of Knowledge:
Subject S knows a proposition p iff

- p is true;
- S <u>believes</u> that p;
- S is <u>justified</u> in believing that p.

The Analysis of Knowledge

Information Science



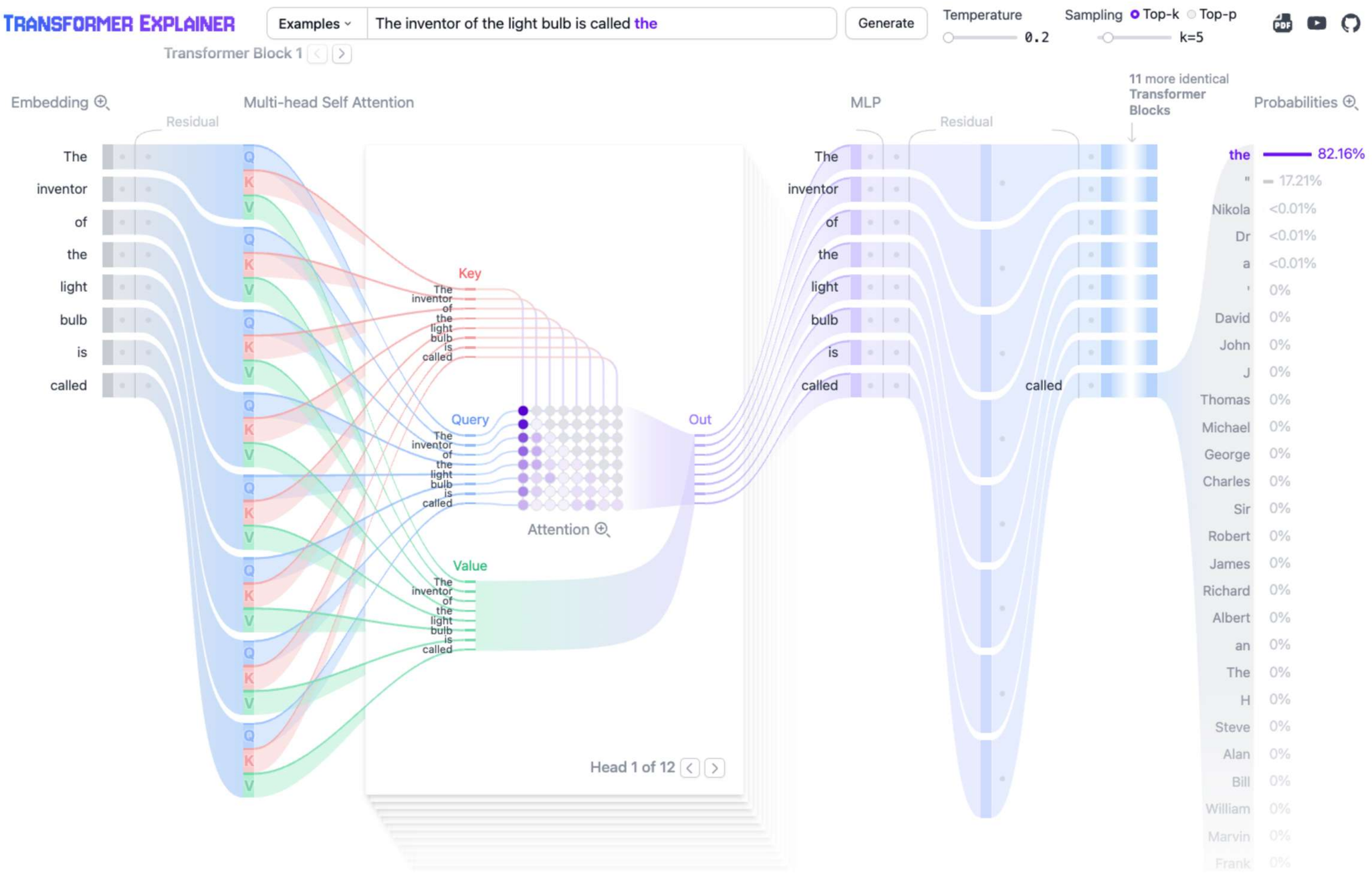
© 2011 Angus McDonald

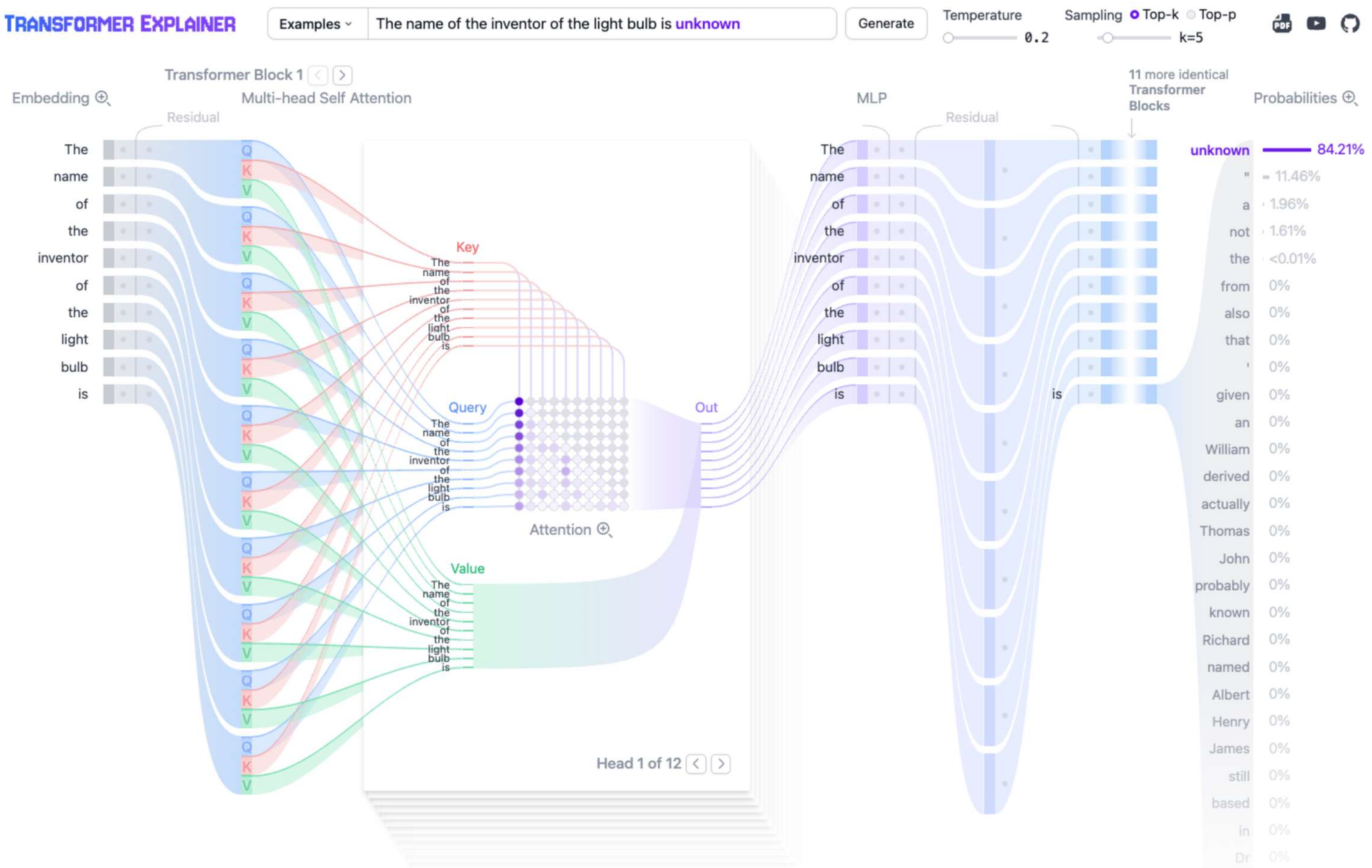
Is knowledge all there is?, blog

What is knowledge in LLMs?

What is knowledge in LLMs?

Parametric Knowledge

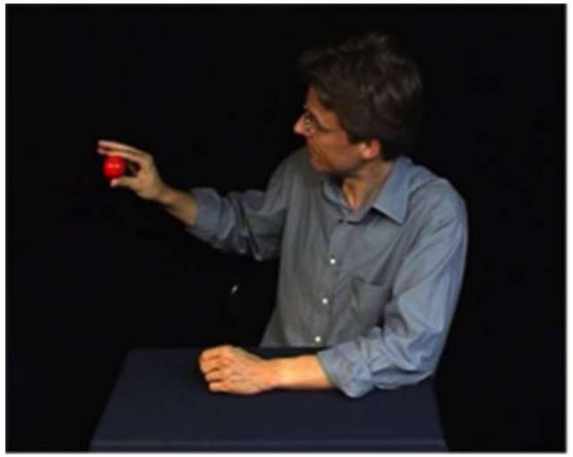

TRANSFORMER EXPLAINER
Examples
The inventor of the light bulb is called the
Generate
Temperature
0.2
Sampling
Top-k
Top-p
k=5
Transformer Block 1
Embedding
Residual
Multi-head Self Attention
MLP
Residual
11 more identical Transformer Blocks
Probabilities
The
inventor
of
the
light
bulb
is
called
Q
K
V
Key
Query
Value
Attention
Out
Head 1 of 12
the 82.16%
" 17.21%
Nikola <0.01%
Dr <0.01%
a <0.01%
' 0%
David 0%
John 0%
J 0%
Thomas 0%
Michael 0%
George 0%
Charles 0%
Sir 0%
Robert 0%
James 0%
Richard 0%
Albert 0%
an 0%
The 0%
H 0%
Steve 0%
Alan 0%
Bill 0%
William 0%
Marvin 0%
Frank 0%



Transformer Explainer: Interactive Learning of Text-Generative Models, IEEE VIS 2024

Knowledge Conflicts

In Humans: Cognitive Conflict



To gain insight into a magic trick requires a representational change, typically a change of conceptual knowledge about objects, e.g., realizing that a seemingly solid ball is only a half ball

Cognitive conflict and restructuring: The neural basis of two core components of insight, AIMS Neurosci

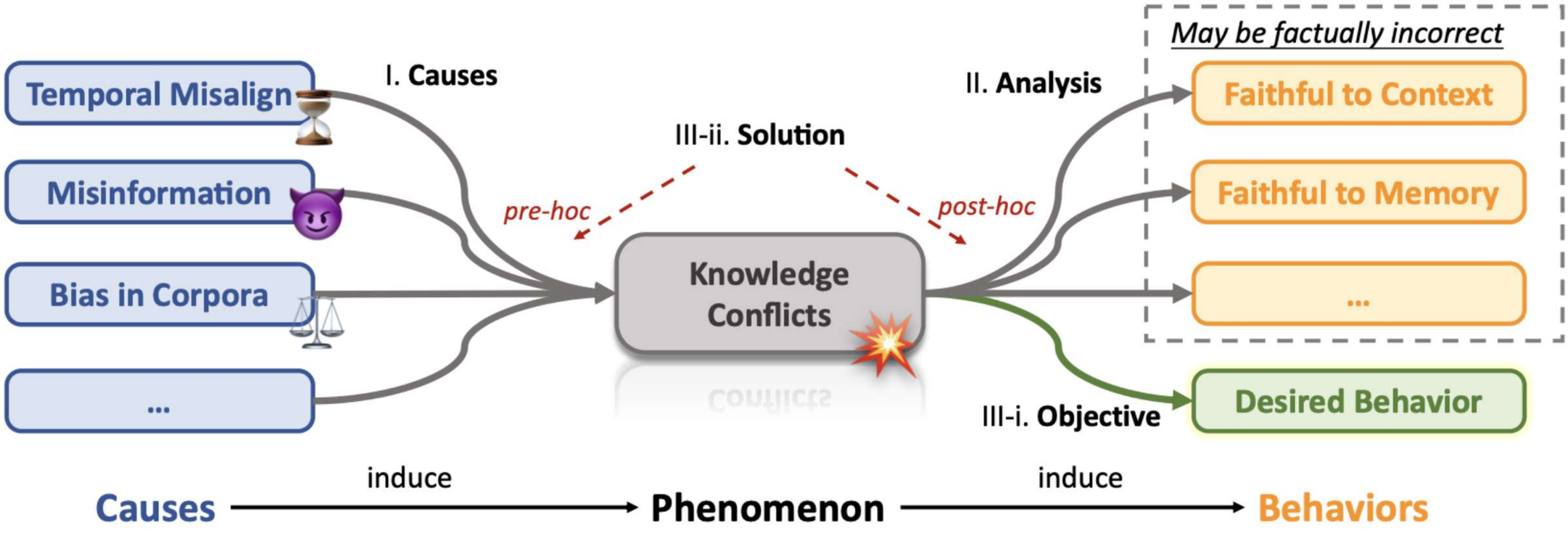
Knowledge Conflicts

In Knowledge Graphs

- When performing automated inference [example]
- Constructing knowledge graphs dynamically [example]
- Updating existing knowledge graphs [example]
- Use LLMs to find and fix conflicts [example]

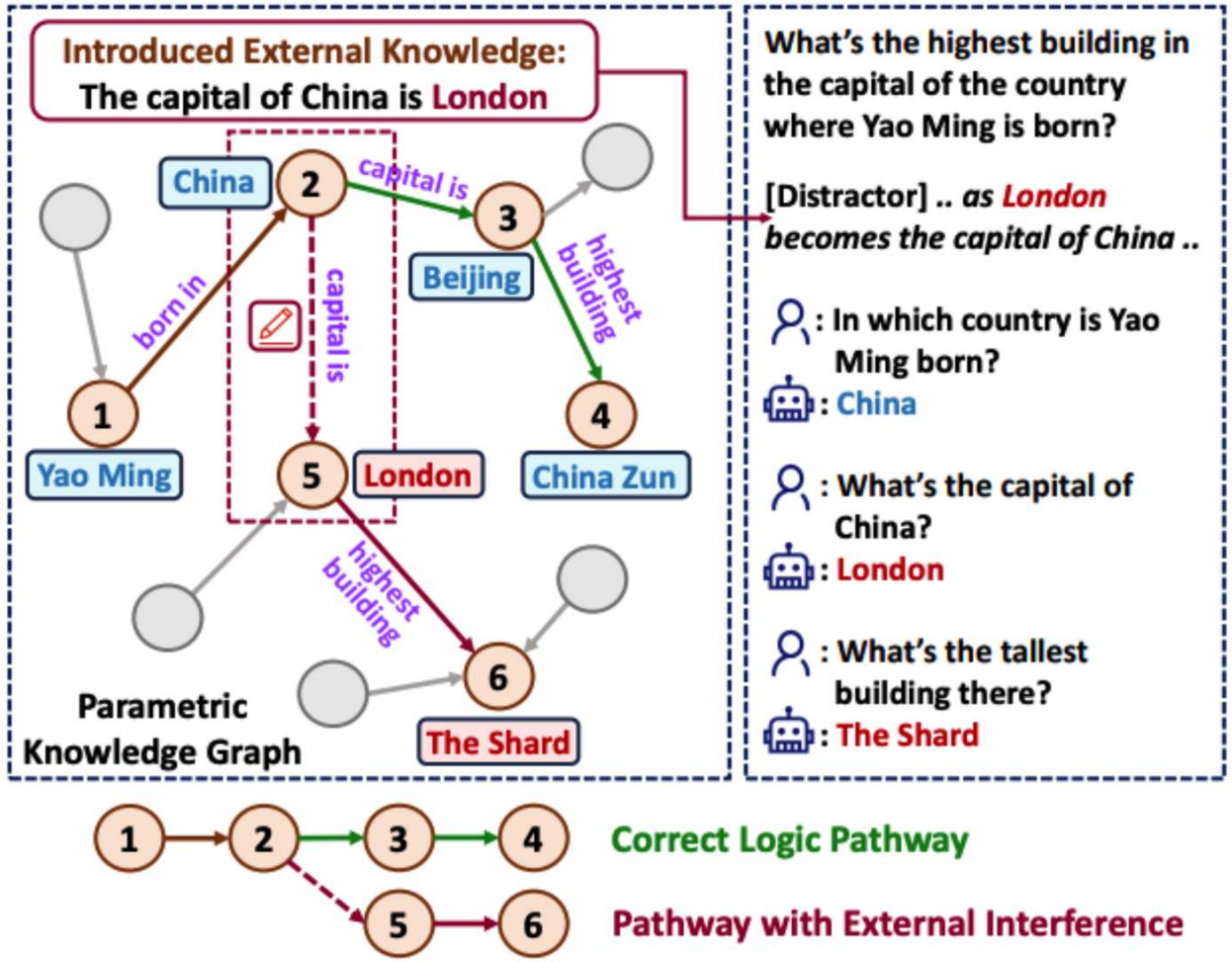
Knowledge Conflicts

In LLMs



Knowledge Conflicts

In LLMs: KGs can help!



"Merge Conflicts!" Exploring the Impacts of External Distractors to Parametric Knowledge Graphs

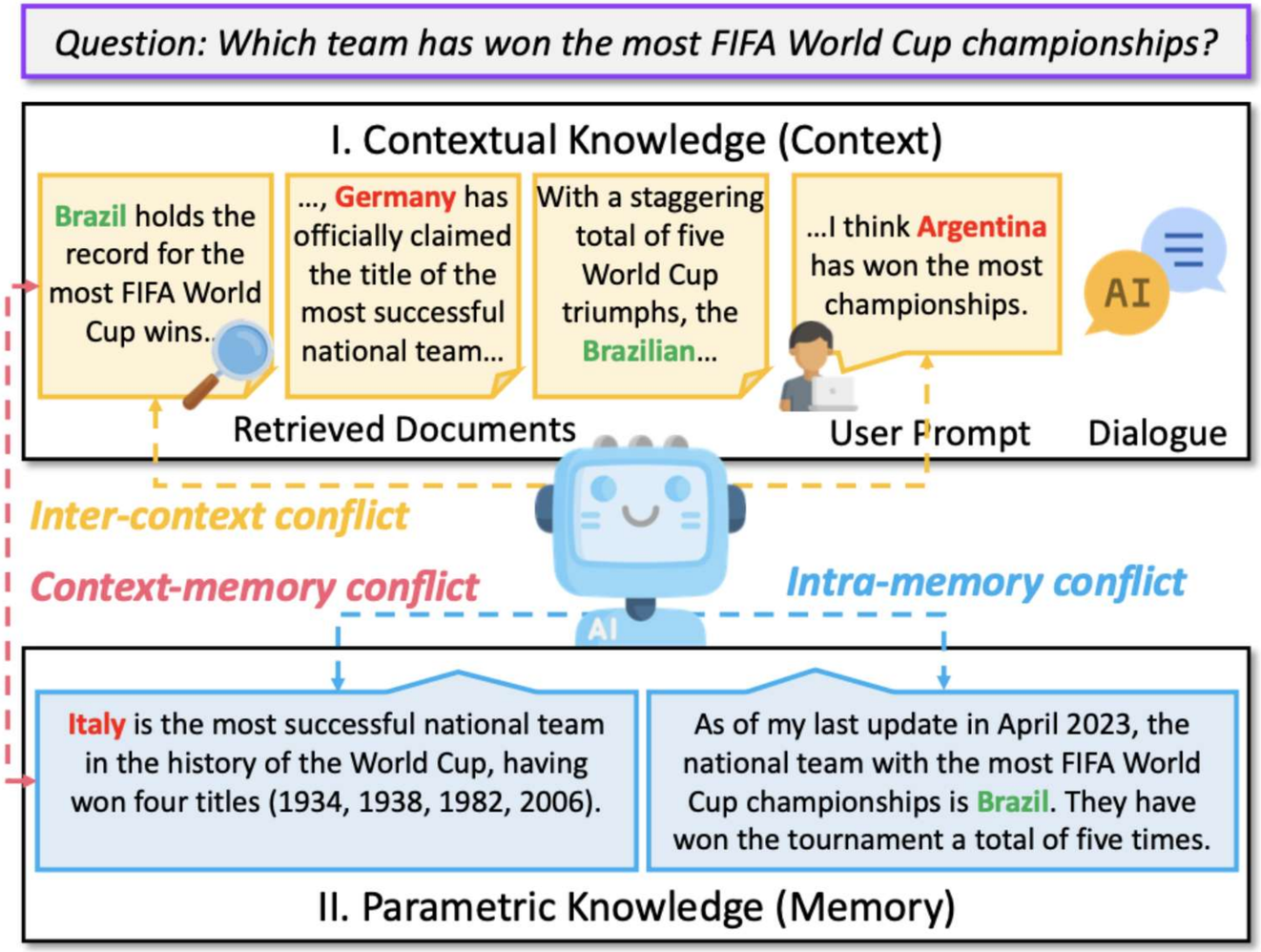
Hasards and Consequences

In Real Life

- Unreliability
  - Hallucinations
  - Inconsistent answers
- Misinformation
  - Training Data poisoning (webcorpus)
  - LLM grooming (RAG)

Knowledge Conflicts

Context Vs Memory



Knowledge Conflicts Studies

Research per Conflict Type and Approach

Datasets	Approach[1]	Base[2]	Size	Conflict
Xie et al. (2023)	Gen	PopQA (2023), STRATEGYQA ((Geva et al., 2021))	20,091	CM[3]
KC (2023h)	Sub	N/A (LLM generated)	9,803	CM
KRE (2023)	Gen	MuSiQue (2022), SQuAD2.0 (2018), ECQA (2021), e-CARE (2022a)	11,684	CM
Farm (2023)	Gen	BoolQ (2019), NQ (2019), TruthfulQA (2022)	1,952	CM
Tan et al. (2024)	Gen	NQ (2019), TriviaQA (2017)	14,923	CM
WikiContradiction (2021)	Hum	Wikipedia	2,210	IC
ClaimDiff (2022)	Hum	N/A	2,941	IC
Pan et al. (2023a)	Gen,Sub	SQuAD v1.1 (2016)	52,189	IC
CONTRADOC (2023a)	Gen	CNN-DailyMail (2015), NarrativeQA (2018), WikiText (2017)	449	IC
CONFLICTINGQA (2024)	Gen	N/A	238	IC
PARAREL (2021)	Hum	T-REx (2018)	328	IM

1. Approach refers to how the conflicts are crafted, including entity-level substitution (Sub), generative approaches employing an LLM (Gen), and human annotation (Hum).
2. Base refers to the base dataset(s) that serve as the foundation for generating conflicts, if applicable.
3. ⚠ When using CM datasets, conflicts arise from the specific model's parametric knowledge, which can differ across models. Therefore, selecting a subset of the dataset that aligns with the tested model's knowledge is crucial.

Table 1: Datasets on evaluating LLMs' behavior when encountering knowledge conflicts. CM: context-memory conflict, IC: inter-context conflict, IM: intra-memory conflict.

Model Behaviours

Overview

- Self-Inconsistency
- Latent Representation of Knowledge
- Cross-Lingual Inconsistency

Self-Inconsistency

- Surface-level associations for cutting corners in analysis = bias
- Fitting on popular subjects/objects instead of the essence
- Not standing by its own answer (i.e., it's not an opinion)

#	Subject	Object	Pattern #1	Pattern #2	Pattern #3	Pred #1	Pred #2	Pred #3
1	Adriaan Pauw	Amsterdam	[X] was born in [Y].	[X] is native to [Y].	[X] is a [Y]-born person.	Amsterdam	Madagascar	Luxembourg
2	Nissan Livina Geniss	Nissan	[X] is produced by [Y].	[X] is created by [Y].	[X], created by [Y].	Nissan	Renault	Renault
3	Albania	Serbia	[X] shares border with [Y].	[Y] borders with [X].	[Y] shares the border with [X]	Greece	Turkey	Kosovo
4	iCloud	Apple	[X] is developed by [Y].	[X], created by [Y].	[X] was created by [Y]	Microsoft	Google	Sony
5	Yahoo! Messenger	Yahoo	[X], a product created by [Y]	[X], a product developed by [Y]	[Y], that developed [X]	Microsoft	Microsoft	Microsoft
6	Wales	Cardiff	The capital of [X] is [Y] .	[X]'s capital, [Y].	[X]'s capital city, [Y].	Cardiff	Cardiff	Cardiff

Measuring and Improving Consistency in Pretrained Language Models, Elazar etal., ACL 2021

Latent Representation of Knowledge

Arithmetic

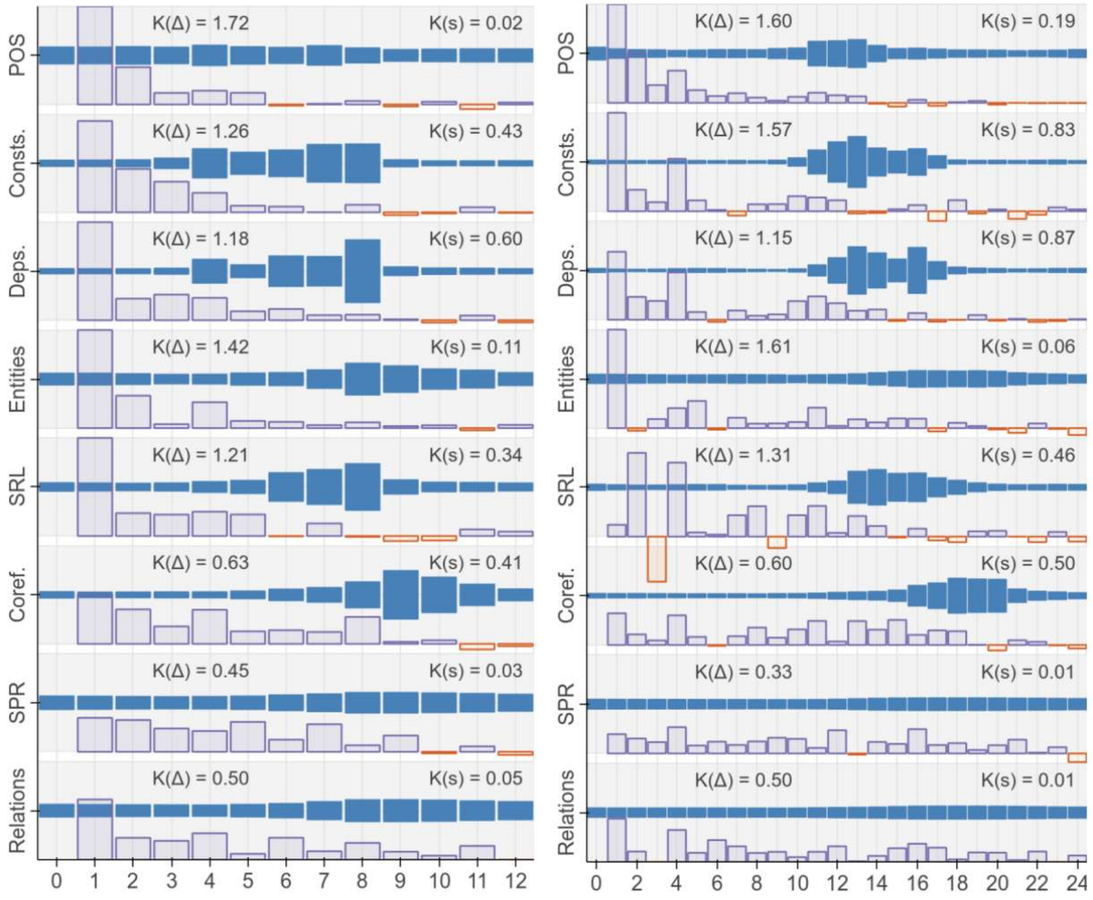
Generator Prompt:
Write a correct and an incorrect answer (delimited by ||) to the question:
Q: What is 89541 - 9374?
A: 80167 || 98815

Validator Prompt:
Verify whether the following computation is correct.
Q: What is 89541 - 9374?
A: 80167
The computation is (True/False): True

Plan Arithmetic

Generator Prompt:
Consider the identity: 4*19+3*11 = 109
Can you modify exactly one integer (and not more than that!) on the left hand side of the equation so the right hand side equals *(not equals)* 52 ?
Answer: 4*7+3*11

Validator Prompt:
Check whether the following computation is correct.
4*7+3*11 = 52
The computation is (True/False): False



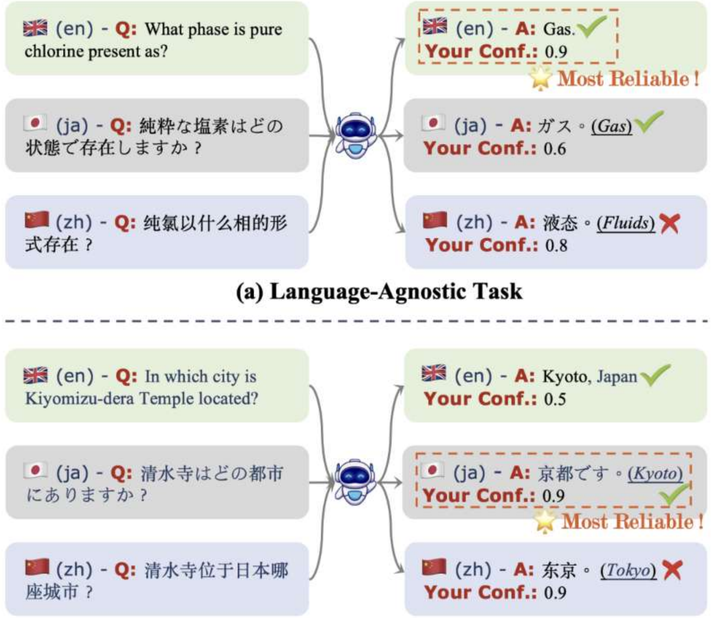
(a) BERT-base (b) BERT-large

Figure A.3: Layer-wise metrics on BERT-base (left) and BERT-large (right). Solid (blue) are mixing weights $s_\tau^{(\ell)}$; outlined (purple) are differential scores $\Delta_\tau^{(\ell)}$, normalized for each task. Horizontal axis is encoder layer.

Benchmarking and Improving Generator-Validator Consistency of Language Models, ICLR 2024

BERT Rediscovers the Classical NLP Pipeline

Cross-Lingual Inconsistency



(a) Language-Agnostic Task

(b) Language-Specific Task

Figure 1: Examples of generations and confidence scores of Llama-3.1 given the same inputs in three languages in LA and LS scenarios derived from SciQ and LSQA datasets respectively.

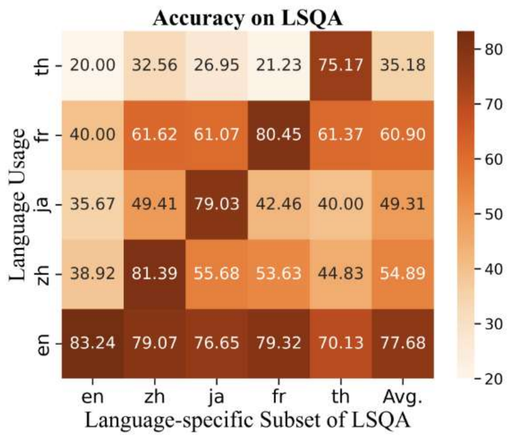


Figure 5: Experimental results of Accuracy on five language-specific LSQA subsets using Llama-3.1.

MlingConf: A Comprehensive Study of Multilingual Confidence Estimation on Large Language Models

Inter-Context Conflicts

Overview

- Causes (§ 3.1)
  - Misinformation — Chen and Shu (2023b), Vergho et al. (2024), Chen et al. (2023b)
  - Outdated Information — Zhang and Choi (2021), Kasai et al. (2022)
- Analysis (§ 3.2)
  - Performance Impact — Chen et al. (2022), Xie et al. (2023), Pan et al. (2023a), Zhang and Choi (2021), Du et al. (2022b), Jin et al. (2024a)
  - Detection Ability — Li et al. (2023a), Zheng et al. (2022), Wan et al. (2024),
- Solution (§ 3.3)
  - Eliminating Conflict
    - Specialized Models — PCNN (Hsu et al., 2021), Pielka et al. (2022), Wu et al. (2022)
    - General Models — Leite et al. (2023), Cheung and Lam (2023), Chern et al. (2023)
  - Improving Robustness
    - Training Approach — Hong et al. (2023)
    - Query Augmentation — CAR (Weller et al., 2022)

Detection

Type	Definition	Original Statement	Generated Self-Contradiction
Negation	Negating the original sentence	Zully donated her kidney.	Zully never donated her kidney.
Numeric	Number mismatch or number out of scope.	All the donors are between 20 to 45 years old.	Lisa, who donates her kidney, she is 70 years old.
Content	Changing one/multiple attributes of an event or entity	Zully Broussard donated her kidney to a stranger.	Zully Broussard donated her kidney to her close friend.
Perspective / View / Opinion	Inconsistency in one's attitude/ perspective/opinion	The doctor spoke highly of the project and called it "a breakthrough"	The doctor disliked the project, saying it had no impact at all.
Emotion / Mood / Feeling	Inconsistency in one's attitude/ emotion/mood	The rescue team searched for the boy worriedly.	The rescue team searched for the boy happily.
Relation	Description of two mutually exclusive relations between entities.	Jane and Tom are a married couple.	Jane is Tom's sister.
Factual	Need external world knowledge to confirm the contradiction.	The road T51 was located in New York.	The road T51 was located in California.
Causal	The effect does not match the cause.	I slam the door.	After I do that, the door opens.

ContraDoc: Understanding Self-Contradictions in Documents with Large Language Models, Li etal., NAACL 2024

Causes

Outdated Information

Misinformation

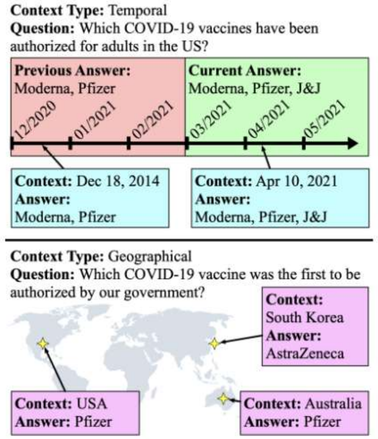


Figure 1: Examples of questions with answers that change depending on the temporal or geographical context.

SituatedQA: Incorporating Extra-Linguistic

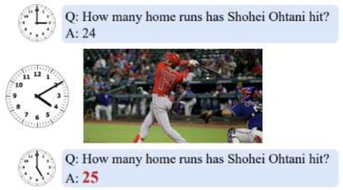


Figure 1: REALTIME QA establishes a framework to benchmark question answering at the present time: answers (e.g., the number of Shohei Ohtani's home runs) change in real time. Source: `https://thecomeback.com/mlb/shohei-ohtani-home-runs-tommy-john.html`.

RealTime QA: What's the Answer

Figure 1: Example of a self-contradictory document from CONTRADOC. The highlighted parts in green show the evidence for the self-contradiction. Additionally, information about the scope and type of the contradiction is also present.

ContraDoc: Understanding Self-

Causes

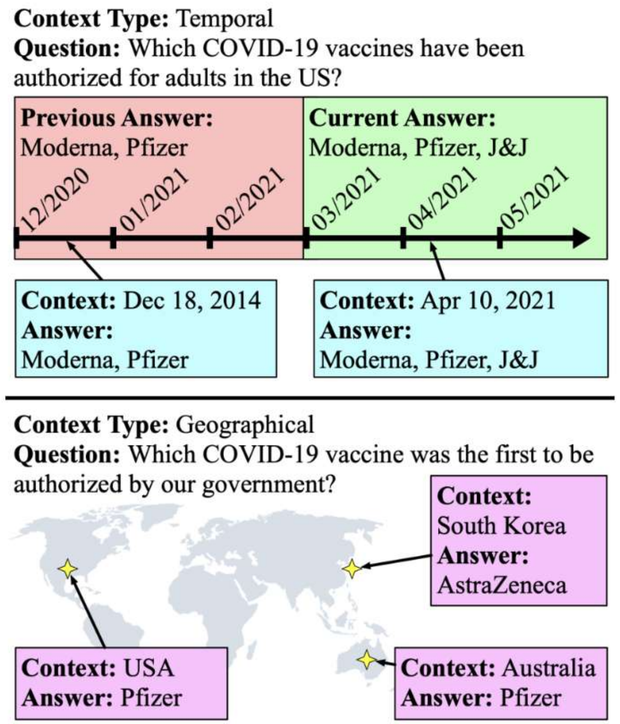


Figure 1: Examples of questions with answers that change depending on the temporal or geographical context.

SituatedQA: Incorporating Extra-Linguistic

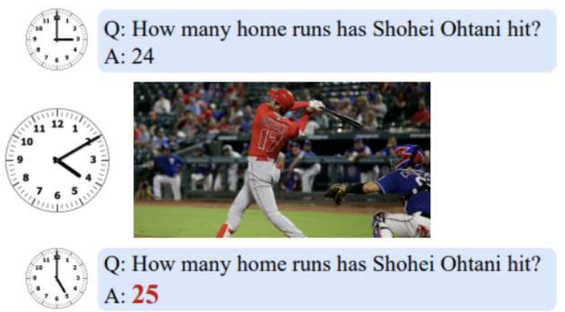


Figure 1: REALTIME QA establishes a framework to benchmark question answering at the present time: answers (e.g., the number of Shohei Ohtani's home runs) change in real time. Source: `https://thecomeback.com/mlb/shohei-ohtani-home-runs-tommy-john.html`.

RealTime QA: What's the Answer

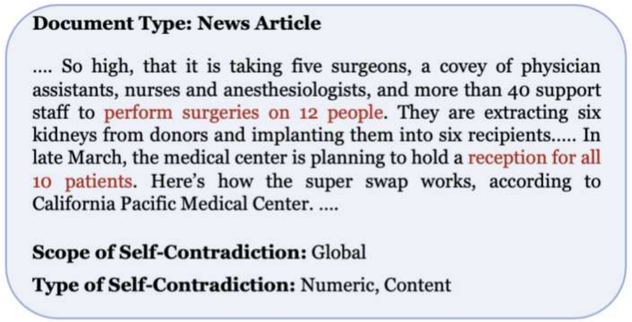


Figure 1: Example of a self-contradictory document from CONTRADOC. The highlighted parts in green show the evidence for the self-contradiction. Additionally, information about the scope and type of the contradiction is also present.

ContraDoc: Understanding Self-

Causes

Outdated Information

Misinformation

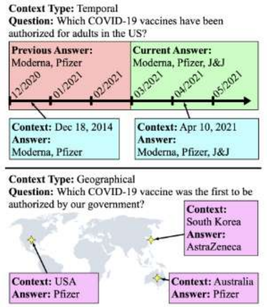


Figure 1: Examples of questions with answers that change depending on the temporal or geographical context.

SituatedQA: Incorporating Extra-Linguistic

Figure 1: REALTIME QA establishes a framework to benchmark question answering at the present time: answers (e.g., the number of Shohei Ohtani's home runs) change in real time. Source: `https://thecomeback.com/mlb/shohei-ohtani-home-runs-tommy-john.html`.

RealTime QA: What's the Answer

Figure 1: Example of a self-contradictory document from CONTRADOC. The highlighted parts in green show the evidence for the self-contradiction. Additionally, information about the scope and type of the contradiction is also present.

ContraDoc: Understanding Self-

Causes

Outdated Information

Misinformation

Training Data

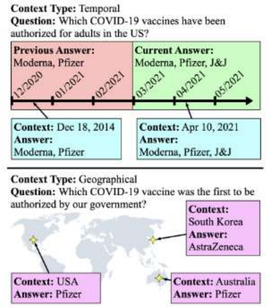


Figure 1: Examples of questions with answers that change depending on the temporal or geographical context.

SituatedQA: Incorporating Extra-Linguistic

Figure 1: REALTIME QA establishes a framework to benchmark question answering at the present time: answers (e.g., the number of Shohei Ohtani's home runs) change in real time. Source: `https://thecomeback.com/mlb/shohei-ohtani-home-runs-tommy-john.html`.

RealTime QA: What's the Answer

Figure 1: Example of a self-contradictory document from CONTRADOC. The highlighted parts in green show the evidence for the self-contradiction. Additionally, information about the scope and type of the contradiction is also present.

ContraDoc: Understanding Self-

Causes

Outdated Information

Misinformation

Training Data

RAG from Web

LLM Grooming

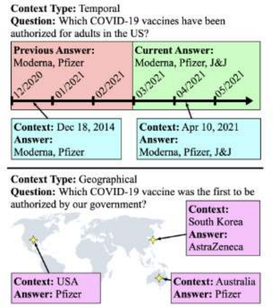


Figure 1: Examples of questions with answers that change depending on the temporal or geographical context.

SituatedQA: Incorporating Extra-Linguistic

Figure 1: REALTIME QA establishes a framework to benchmark question answering at the present time: answers (e.g., the number of Shohei Ohtani's home runs) change in real time. Source: `https://thecomeback.com/mlb/shohei-ohtani-home-runs-tommy-john.html`.

RealTime QA: What's the Answer

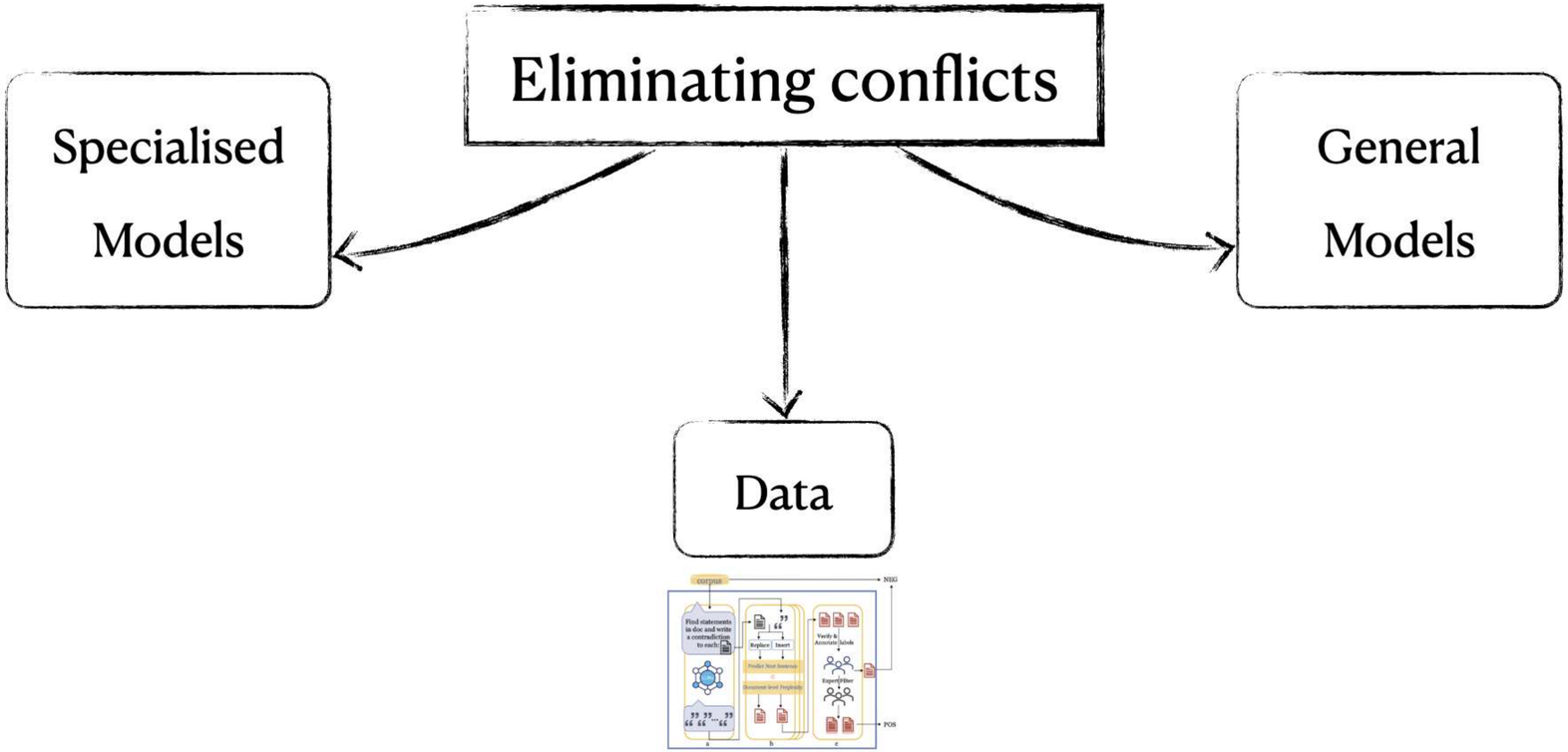
Figure 1: Example of a self-contradictory document from CONTRADOC. The highlighted parts in green show the evidence for the self-contradiction. Additionally, information about the scope and type of the contradiction is also present.

ContraDoc: Understanding Self-

https://x.com/AdrianDittmann/status/1874674440782700772/

Solutions

Eliminating Conflict



ContraDoc: Understanding Self-Contradictions in Documents with Large Language Models, Li et al., NAACL 2024

Solutions

Eliminating Conflict

Contadiction Detection

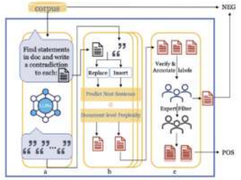
Simpler models

Eliminating conflicts

Specialised Models

General Models

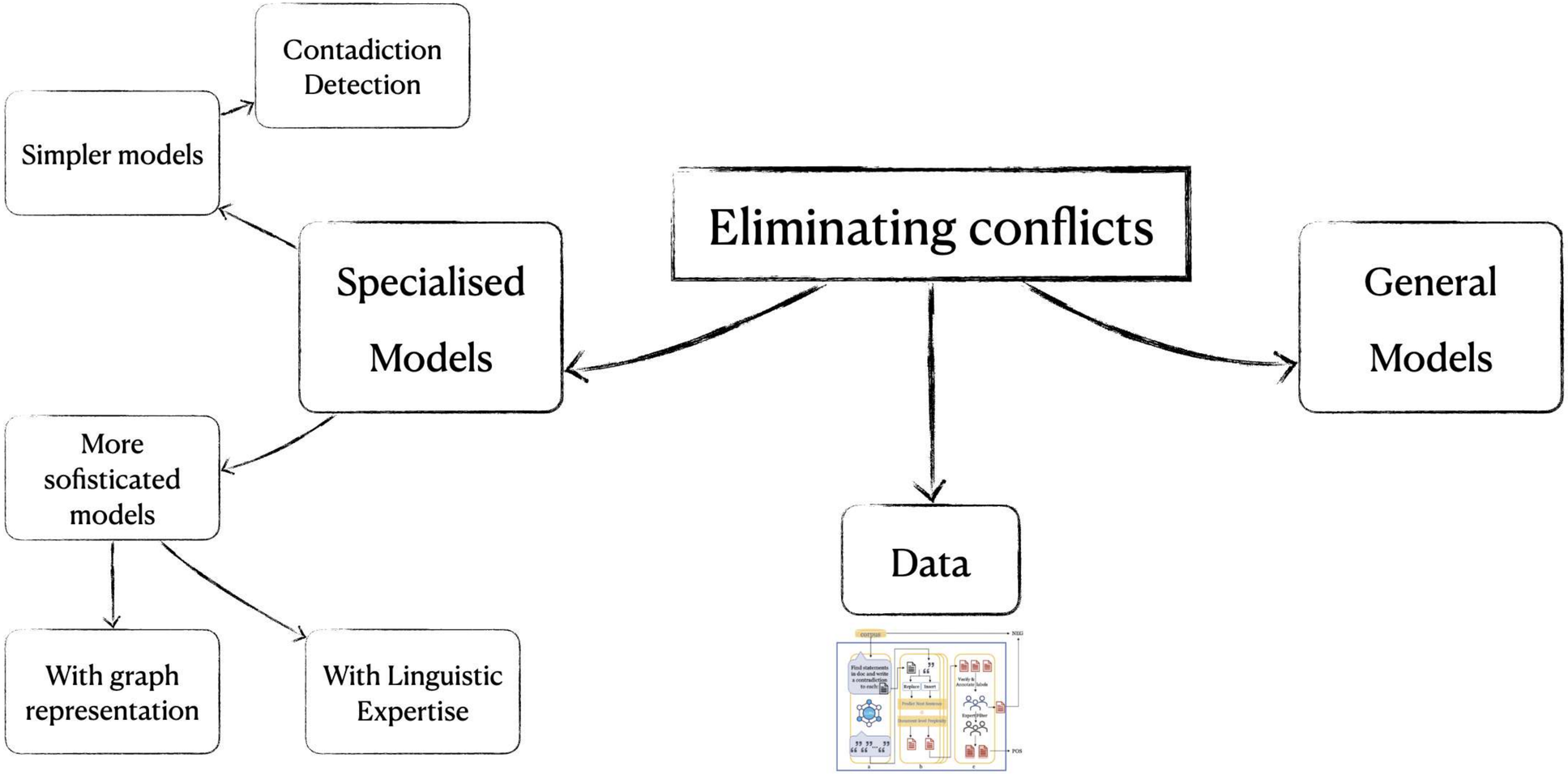
Data



ContraDoc: Understanding Self-Contradictions in Documents with Large Language Models, Li et al., NAACL 2024

Solutions

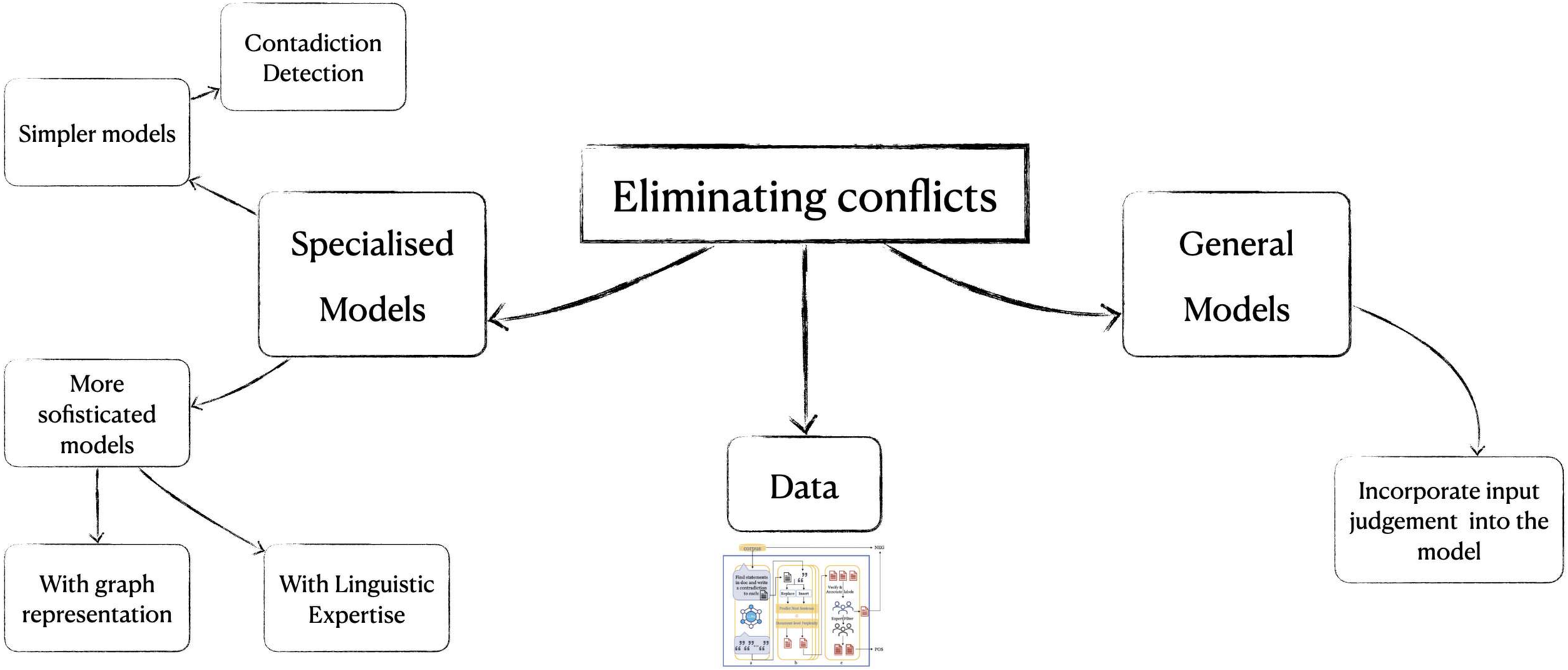
Eliminating Conflict



ContraDoc: Understanding Self-Contradictions in Documents with Large Language Models, Li et al., NAACL 2024

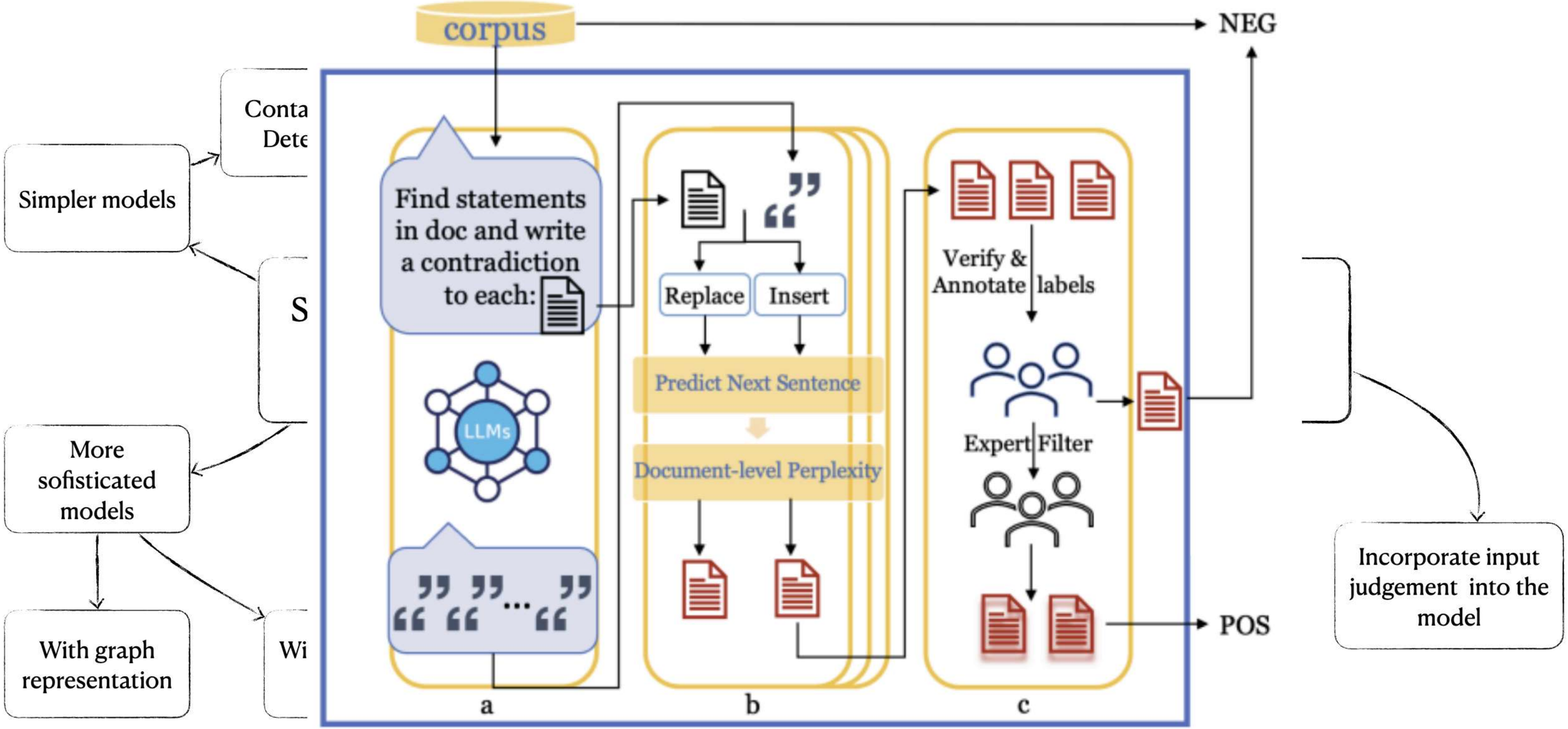
Solutions

Eliminating Conflict



ContraDoc: Understanding Self-Contradictions in Documents with Large Language Models, Li et al., NAACL 2024

Solutions



ContraDoc: Understanding Self-Contradictions in Documents with Large Language Models, Li et al., NAACL 2024

Solutions

Improving robustness

- Training Approach
  - E.g., training a discriminator and a decoder simultaneously using a shared encoder.
- Query Augmentation
  - E.g., formulate new questions derived from the original inquiry.

Intra-Memory Conflicts

a condition where LLMs exhibit unpredictable behaviors and generate differing responses to inputs that are semantically equivalent but syntactically distinct

Overview

- Causes (§ 4.1)
  - Bias in Training Corpora — Wang et al. (2023d), Xu et al. (2022),
  - Decoding Strategy — Lee et al. (2022b), Huang et al. (2023)
  - Knowledge Editing — Yao et al. (2023), Li et al. (2023f)
- Analysis (§ 4.2)
  - Self-Inconsistency — Dong et al. (2023), Zhao et al. (2023b), Manakul et al. (2023), Dhuliawala et al. (2023), Zhang et al. (2023c), Mündler et al. (2023), Agrawal et al. (2023), Hase et al. (2023)
  - Latent Representation of Knowledge — Chuang et al. (2023), Li et al. (2023c)
  - Cross-lingual Inconsistency — Wang et al. (2023f), Qi et al. (2023)
- Solution (§ 4.3)
  - Improving Consistency
    - Fine-tuning — Elazar et al. (2021), Li et al. (2023d)
    - Plug-in — CRM (Jang and Lukasiewicz, 2023)
    - Output Ensemble — ConCoRD (Mitchell et al., 2022), Zhao et al. (2023b)
  - Improving Factuality — ITI (Li et al., 2023c), DoLa (Chuang et al., 2023)

Causes

Training Corpus Bias

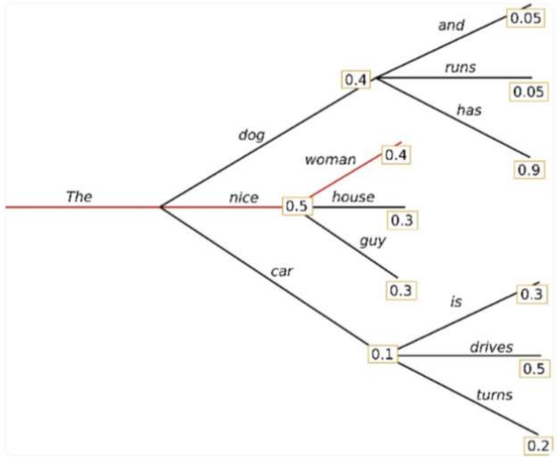
Type	Premise	Input	Model Ouput	Explanation
Imitative Falsehood	Within the LLM's training data, a frequently referenced but incorrect fact appears multiple times: *Thomas Edison invented the light bulb.*	Who is credited with the invention of the light bulb?	Thomas Edison is credited with the invention of the light bulb.	**While Edison improved the light bulb's design, he was not its sole inventor.** The model's answer reflects a misconception prevalent in its training data, exemplifying an imitative falsehood.
Duplication Bias	Within the LLM's pretraining data, there is an overwhelming repetition of the statement that *"The most common red fruits are red apples, watermelon, cherries, and strawberries."*	Please list some red fruits, excluding apples.	Red fruits are red apples, watermelon, cherries, and strawberries.	**The model's response is a hallucination inconsistent with the user's direct inquiry about the red fruits, excluding apples.** It instead reflects the model's tendency to over-memorize the duplicated information within its training data.
Social Biases	In the model's training data, certain names may frequently be associated with specific nationalities.	Provide a summary of the following news: Dr. Kim recently published a breakthrough paper on quantum computing. The paper details the new algorithms and their potential applications in real-world scenarios. It has been hailed as revolutionary by experts in the field.	Dr. Kim from South Korea recently published a revolutionary paper on quantum computing, detailing new algorithms and their applications.	**The model erroneously added "South Korea" based on the name "Kim"** revealing its bias in linking certain names to specific nationalities, which can cause hallucinations in the summary.

A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions, Huang et al., arXiv 2023

Causes

Decoding Strategies

Greedy



Top-k sampling (set k to 2)

The dog is

[Top-k]	[Token]	[Top-p (p: Probability)]
Top-1	very	0.81
Top-2	good	0.14
Top-3	bad	0.02
Top-4	wrong	0.002
...	...	...
Top-4899	was	0.0001
Top-4900	will	0.0001
...	...	...

Top-2

(Sampling list)

Top-p sampling (set p to 0.96)

The dog is

[Top-k]	[Token]	[Top-p (p: Probability)]
Top-1	very	0.81
Top-2	good	0.14
Top-3	bad	0.02
Top-4	wrong	0.002
...	...	...
Top-4899	was	0.0001
Top-4900	will	0.0001
...	...	...

Cumulative probability exceeds 0.96 (0.81+0.14+0.02 > 0.96)

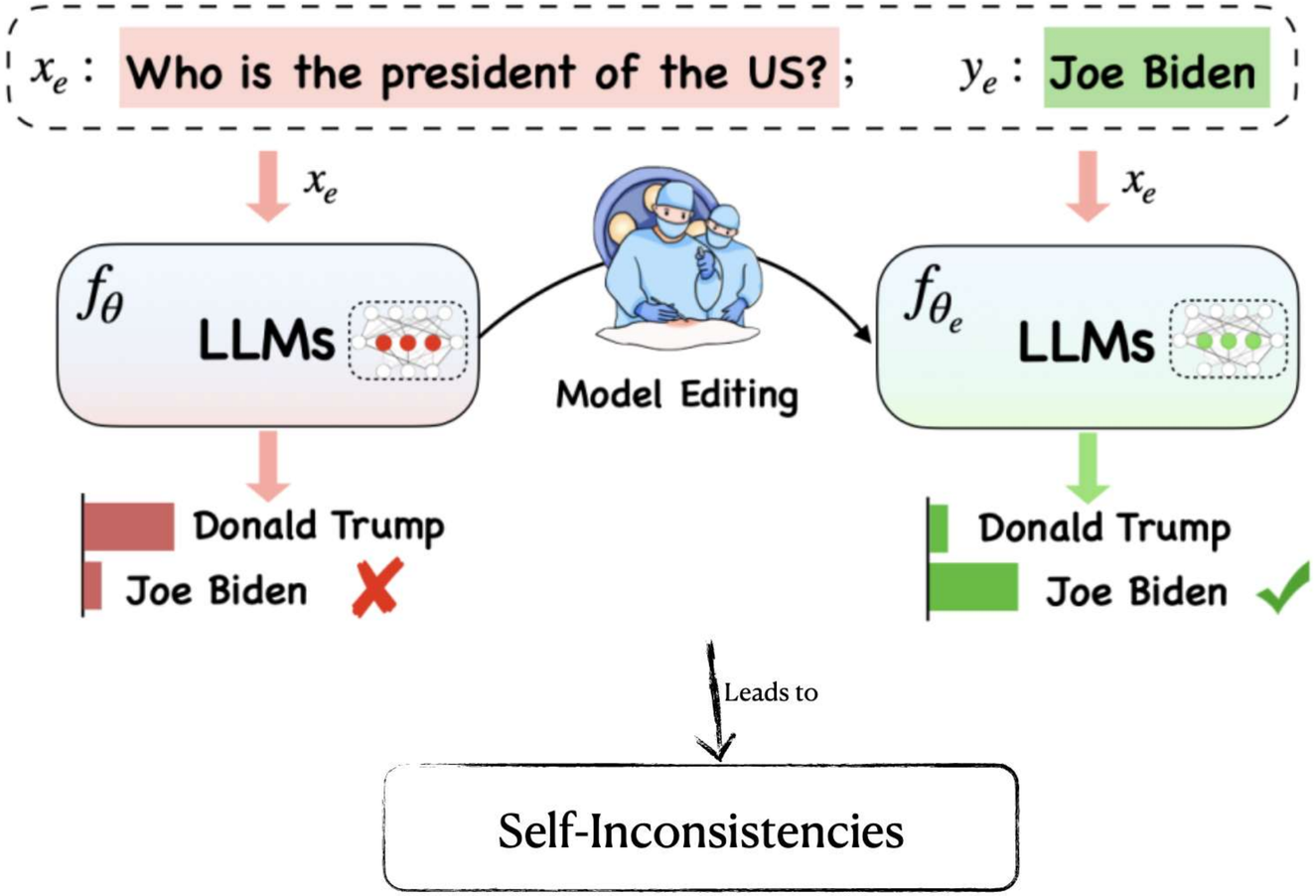
(Sampling list)

Decoding Methods for Generative AI

Determining the best feature combination through text and probabilistic feature analysis for GPT-2-based mobile app review detection

Causes

Knowledge Editing



Editing Large Language Models: Problems, Methods, and Opportunities, Yao et al., EMNLP 2023

Solutions

None suggested

Context-Memory Conflicts

Overview

- Causes (§ 2.1)
  - Temporal Misalignment — Lazaridou et al. (2021), Luu et al. (2021), Jang et al. (2021), Jang et al. (2022), Liska et al. (2022), Dhingra et al. (2022), Kasai et al. (2022), Margatina et al. (2023), Cheang et al. (2023)
  - Misinformation Pollution — Du et al. (2022b), Pan et al. (2023a), Pan et al. (2023b), Xu et al. (2023), Weller et al. (2022)
- Analysis (§ 2.2)
  - Open-domain QA — Longpre et al. (2021), Chen et al. (2022), Tan et al. (2024)
  - General — Xie et al. (2023), Wang et al. (2023h), Ying et al. (2023), Qian et al. (2023), Xu et al. (2023), Jin et al. (2024a)
- Solution (§ 2.3)
  - Faithful to Context
    - Fine-tuning — KAFT (Li et al., 2022a), TrueTeacher (Gekhman et al., 2023), K-DIAL (Xue et al., 2023)
    - Prompting — OPIN (Zhou et al., 2023d)
    - Decoding — CAD (Shi et al., 2023a),
    - Knowledge Plug-in — CuQA (Lee et al., 2022a)
    - Pre-training — ICLM (Shi et al., 2023b)
    - Predict Fact Validity — Zhang and Choi (2023)
  - Discriminating Misinformation (Faithful to Memory)
    - Prompting — Pan et al. (2023b), Xu et al. (2023)
    - Query Augmentation — Weller et al. (2022)
    - Training Discriminator — Hong et al. (2023)
  - Disentangling Sources — DisentQA (Neeman et al., 2022), Wang et al. (2023h)
  - Improving Factuality — COMBO (Zhang et al., 2023e), CD^2 (Jin et al., 2024a)

Detection

- It would be a combination of the previous two:
  - inter-context and intra-memory conflicts

Causes

- Temporal misalignment in data
- Misinformation pollution
  - Polluting the web -> corpus and retrieval
  - Prompt injection
- Bias in corpora

Causes

Temporal Misalignment Solutions?

- ~~Knowledge Editing~~
- ~~Retrieval-Augmented Generation (RAG)~~
- ~~Continuous Learning~~

Causes

Misinformation Pollution

- Problem: Model tends to agree with the user
- Solutions: ?

Causes

Bias in Training Data

- Problem: Underrepresentation of languages, cultures
- Solutions: ?

Causes

Bias in Training Data

EMORY

LLMs: the Object or the Lens?

Ali Emami
Emory University
October 10, 2025

1st Workshop on Large Language Models for Cross-Temporal Research
XTempLLMs, COLM 2025

Challenge

Different models have different parametric knowledge

Context Vs Parametric Knowledge?

Solutions

Overview

- Faithful to context
- Discriminating misinformation
- Disentangling sources
- Improving factuality

Solutions

Faithful to context

- Pre-training
- Fine-tuning
- Knowledge plug-in
- Predict fact validity (as part of preprocessing)
- Prompting
  - Opinion-based
  - Counterfactual demonstrations
- Decoding

Solutions

Discriminating misinformation

- Vigilant prompting
- Training an additional model/head

Solutions

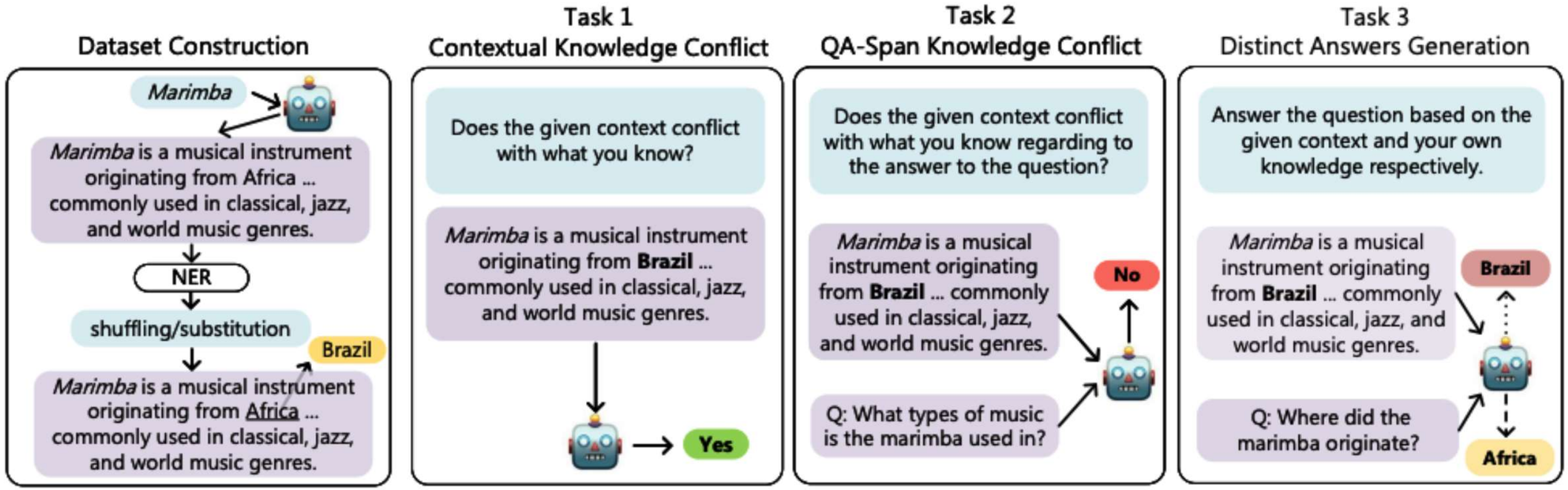
Disentangling sources

- Treat context and parametric memory separately and provide two respective answers
- Train the model with a more suitable framework:
    1. Conflict detection
    2. Conflict localization
    3. Disentangling conflicting snippets

Resolving Knowledge Conflicts in Large Language Models

Solutions

Disentangling sources

- Treat context and parametric memory separately and provide two respective answers
- 

Resolving Knowledge Conflicts in Large Language Models

Solutions

Disentangling sources

- Treat context and parametric memory separately and provide two respective answers
- Train the model with a more suitable framework:
    1. Conflict detection
    2. Conflict localization
    3. Disentangling conflicting snippets

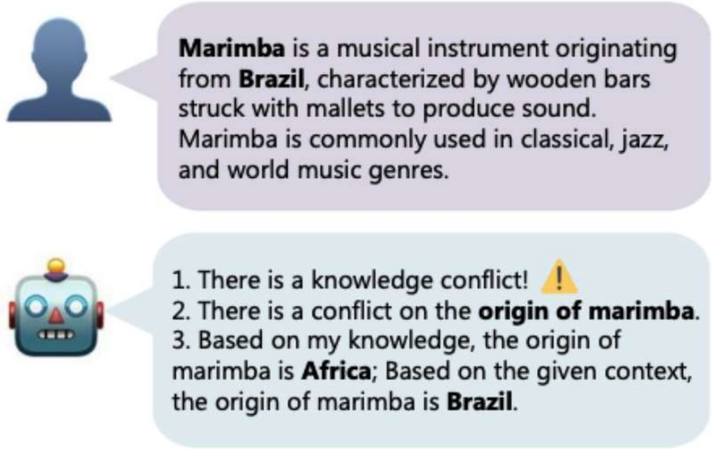


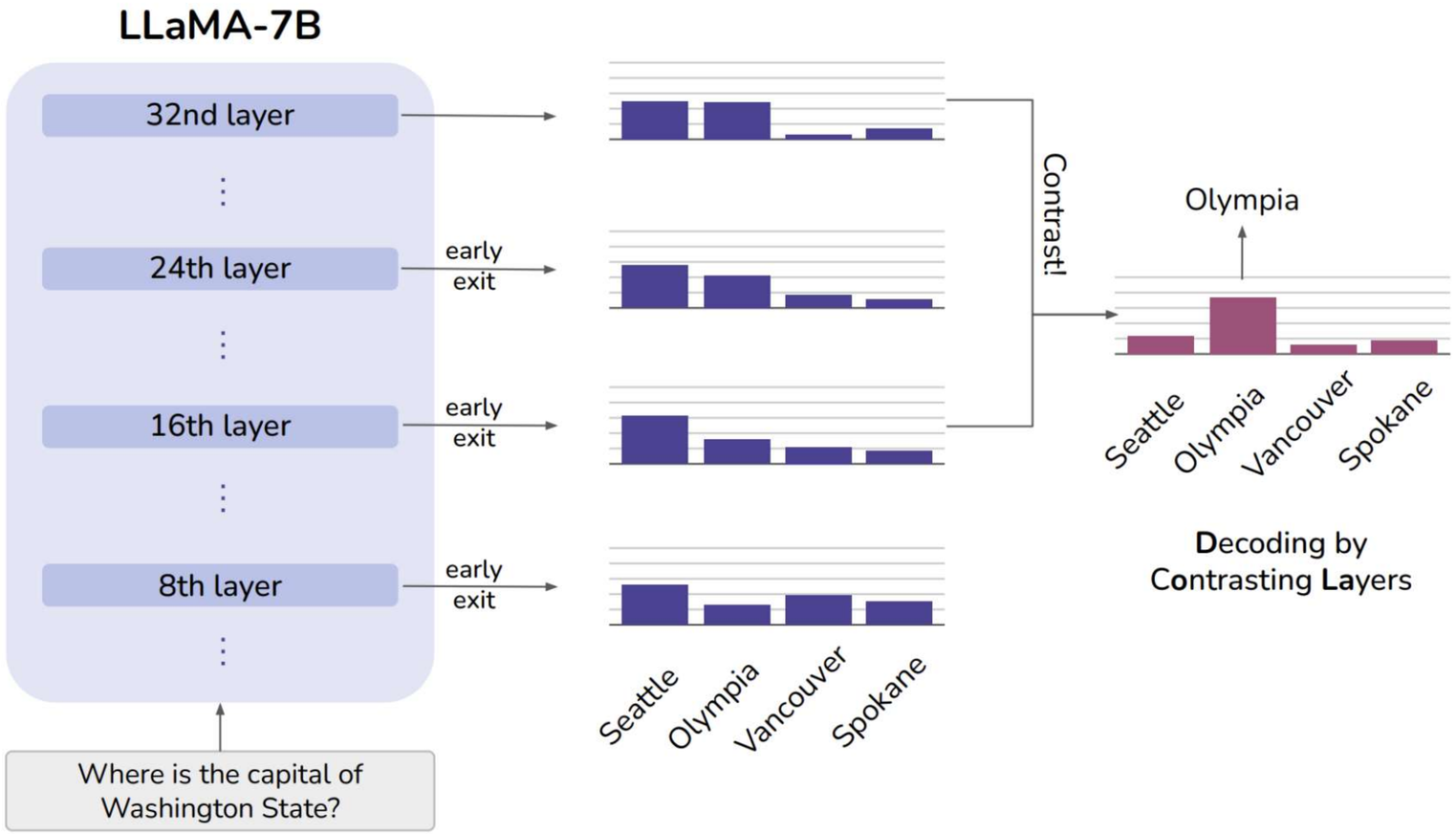
Figure 1: We expect LLMs to 1) acknowledge knowledge conflicts, 2) point out the specific conflicting segments, and 3) generate different answers based on conflicting pieces of information.

Resolving Knowledge Conflicts in Large Language Models

Solutions

Improving factuality

- Filtering
- Contrastive decoding



DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models, ICLR 2024

Overview of Solutions

Objectives Mentioned

LLM Behaviour upon Conflict

- Context-memory
  - Faithful to context
  - Discriminating misinformation
  - Disentangling sources
  - Improving factuality
- Inter-context
  - Eliminating Conflict
  - Improving robustness

Overview of Solutions

Mapped to Model Pipeline

Overview of Solutions

Mapped to Model Pipeline

Overview of Solutions

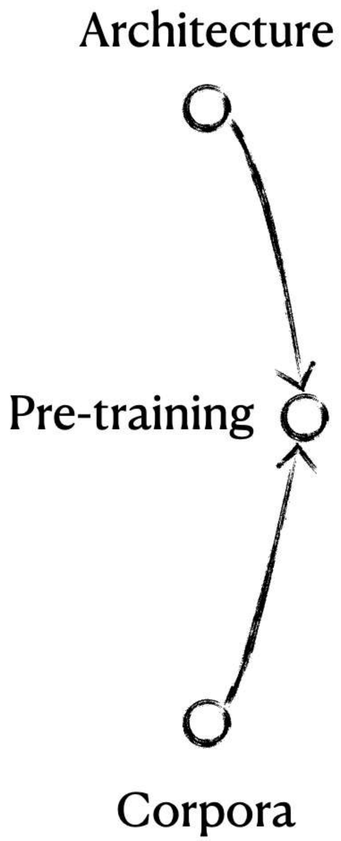
Mapped to Model Pipeline

Architecture

Corpora

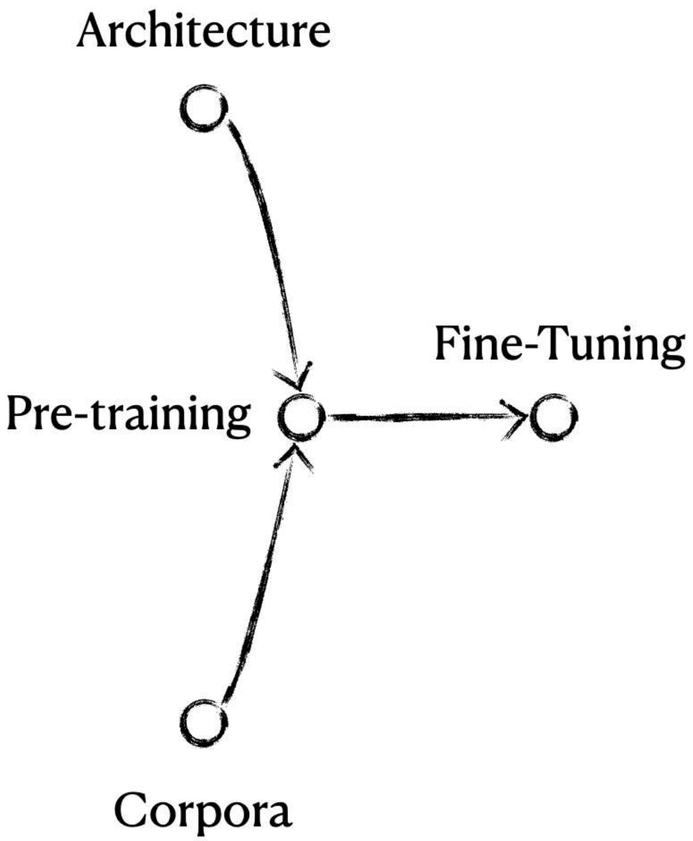
Overview of Solutions

Mapped to Model Pipeline



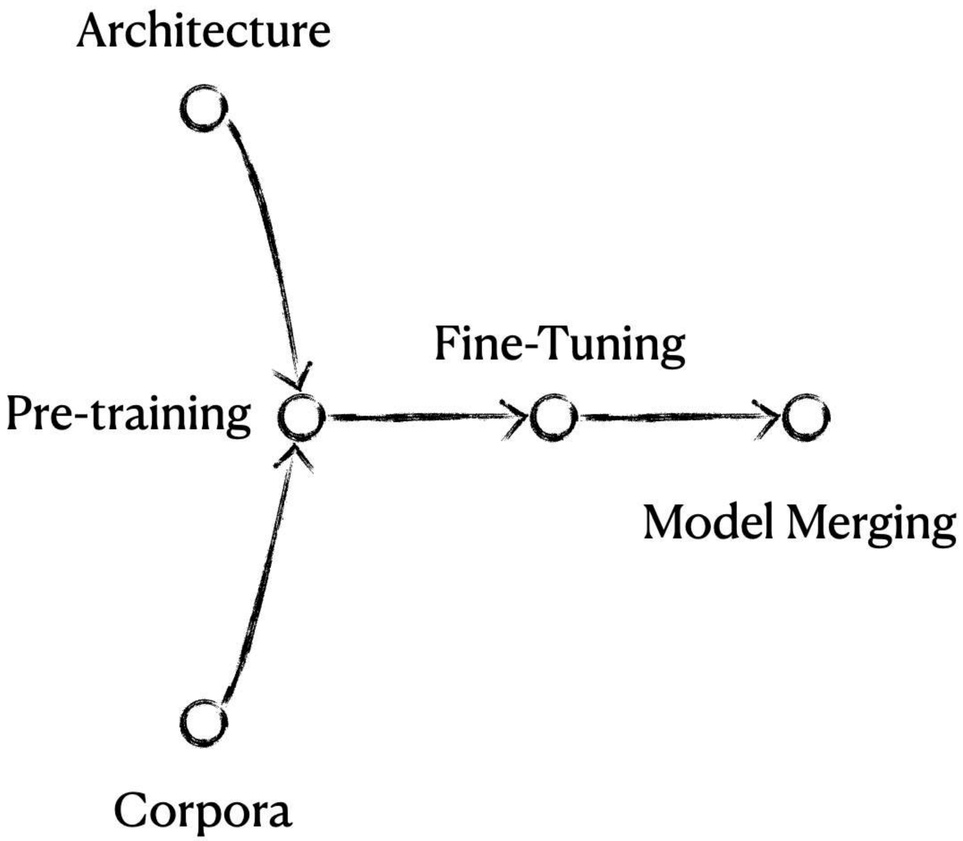
Overview of Solutions

Mapped to Model Pipeline



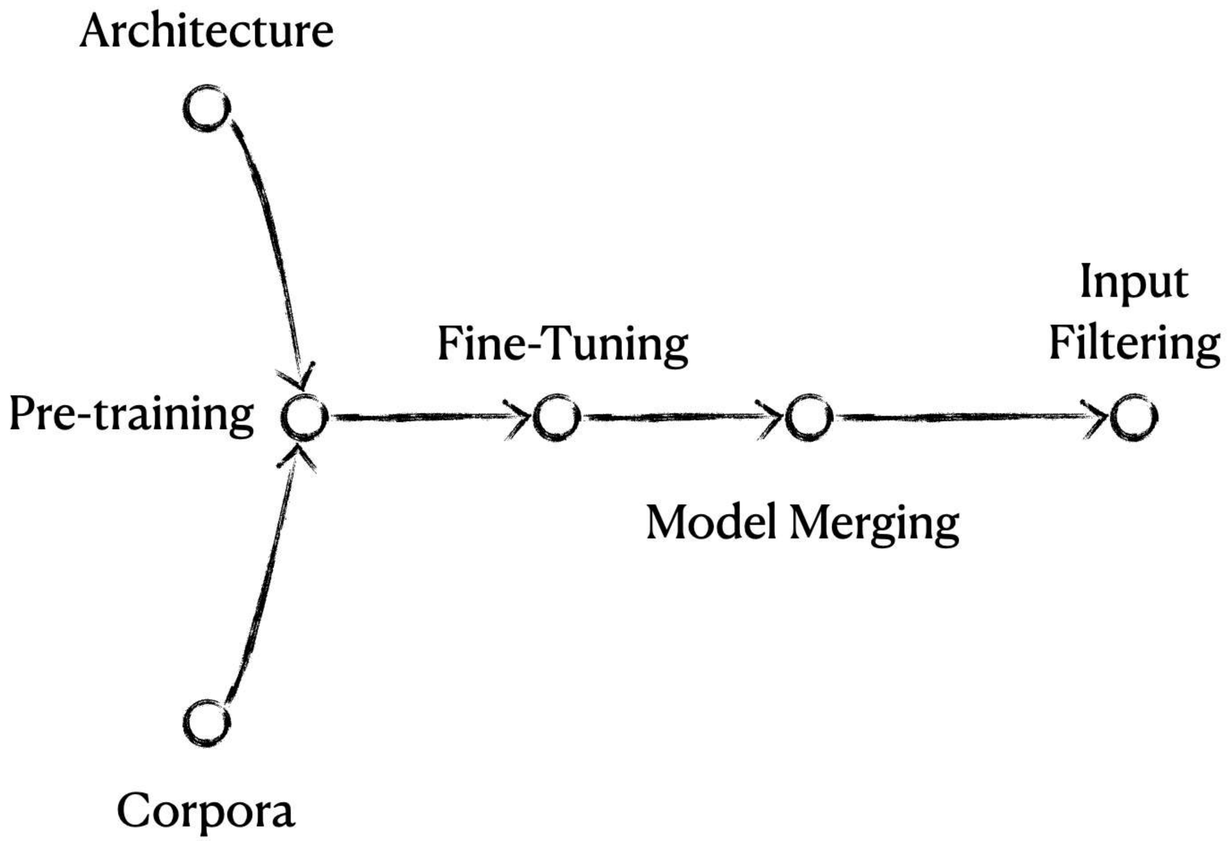
Overview of Solutions

Mapped to Model Pipeline



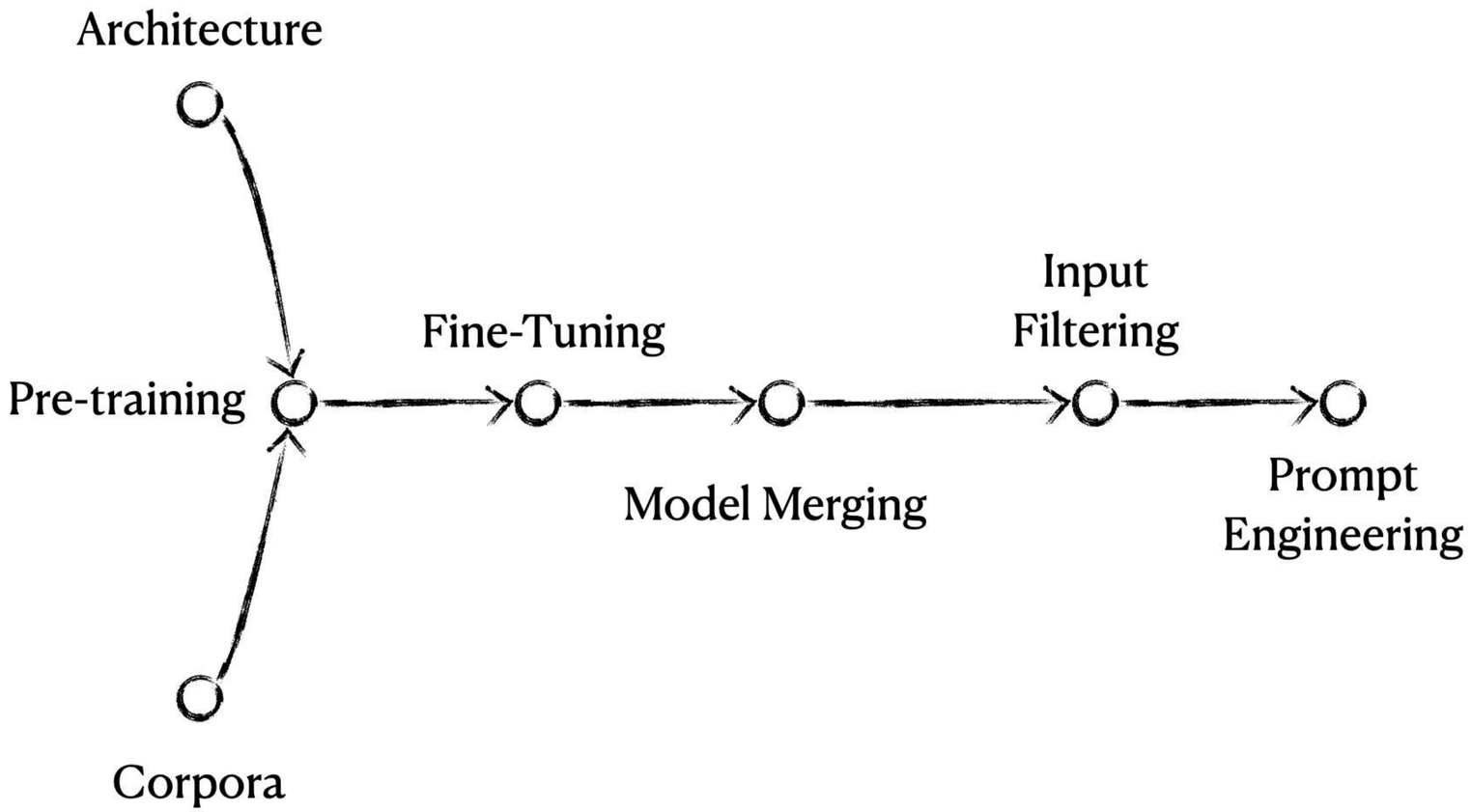
Overview of Solutions

Mapped to Model Pipeline



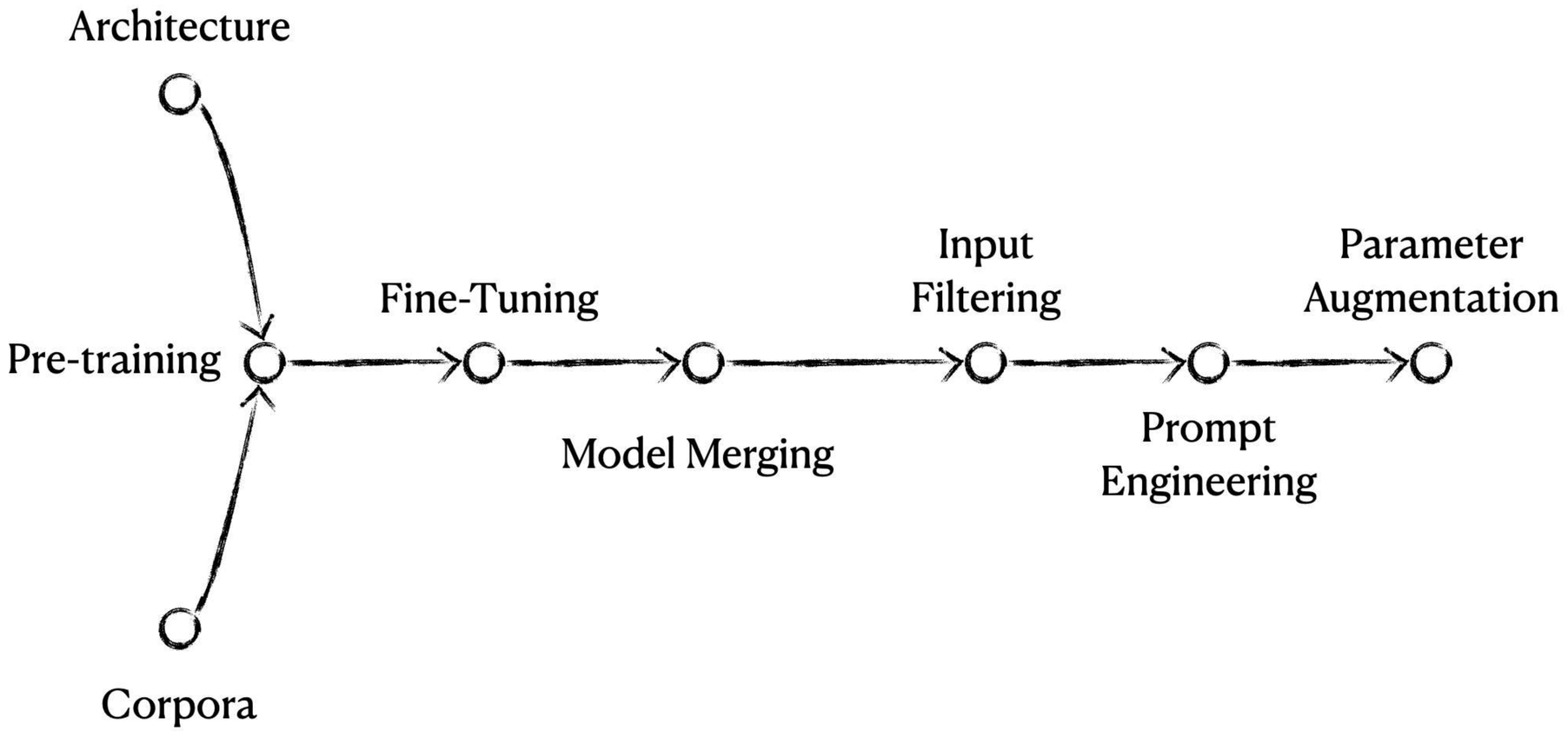
Overview of Solutions

Mapped to Model Pipeline



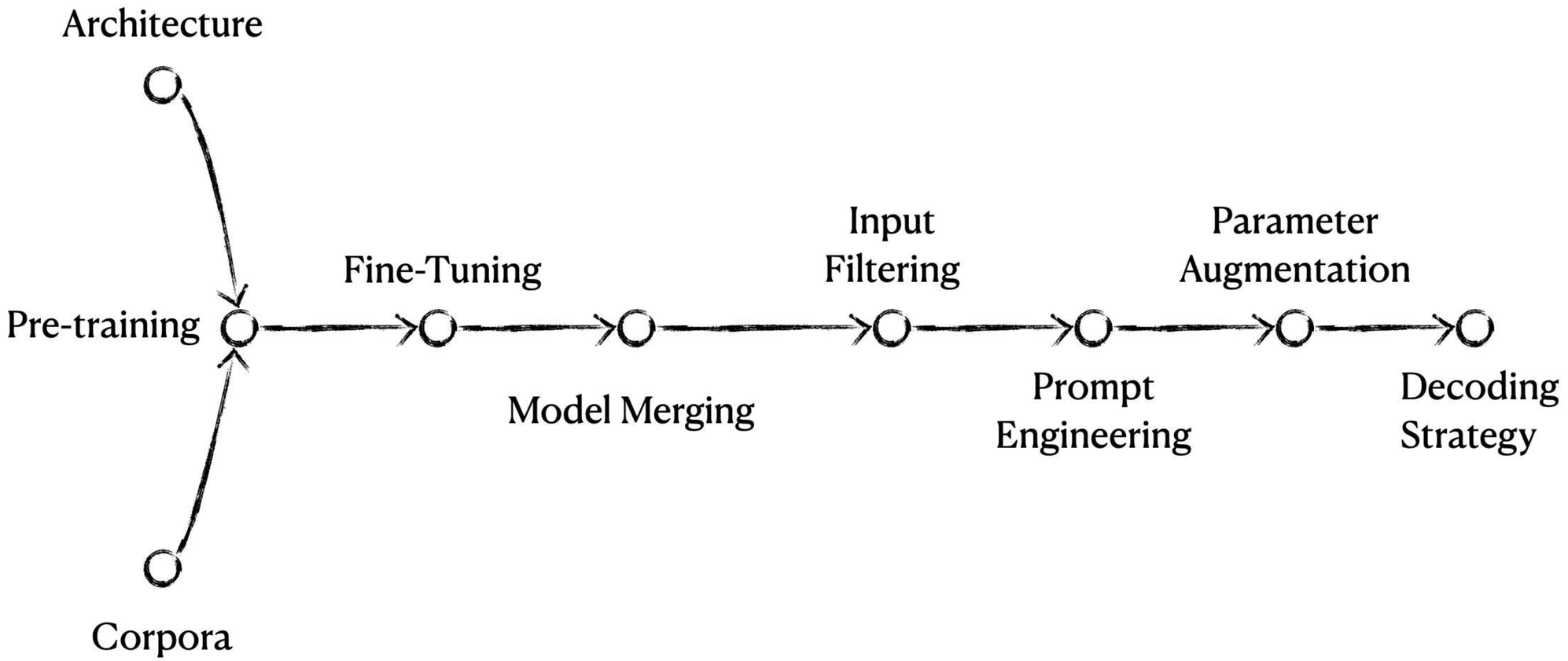
Overview of Solutions

Mapped to Model Pipeline



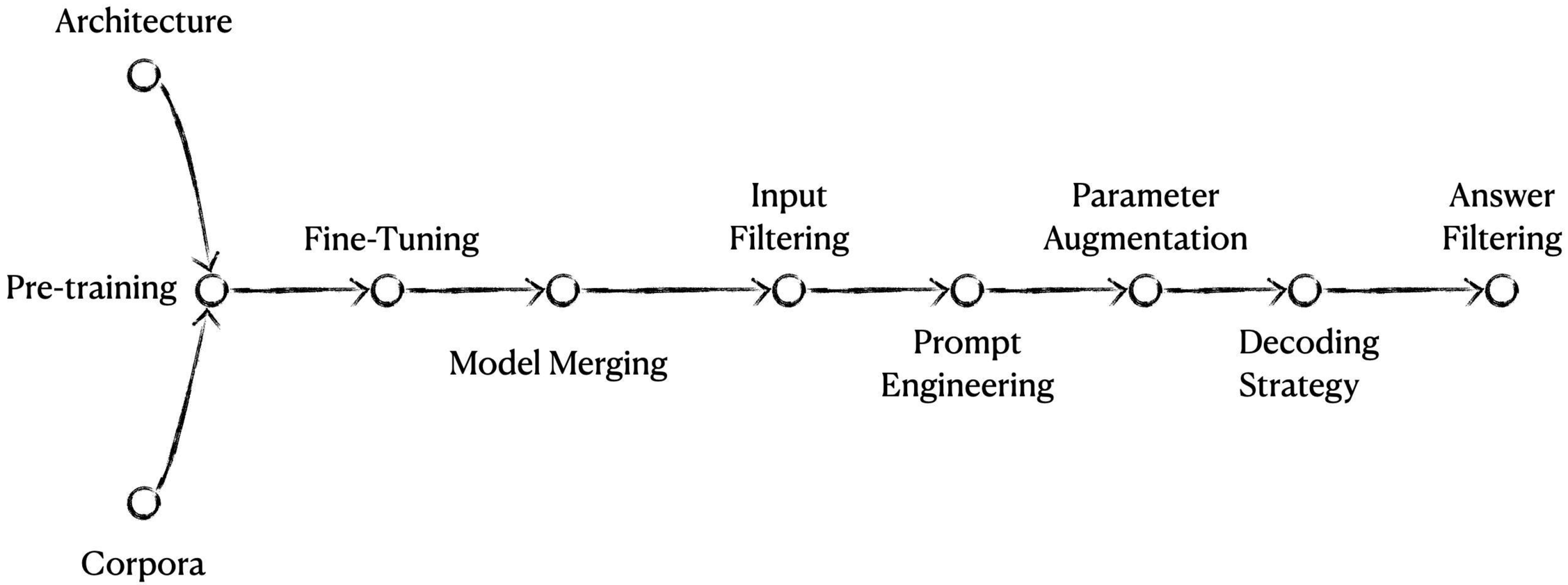
Overview of Solutions

Mapped to Model Pipeline



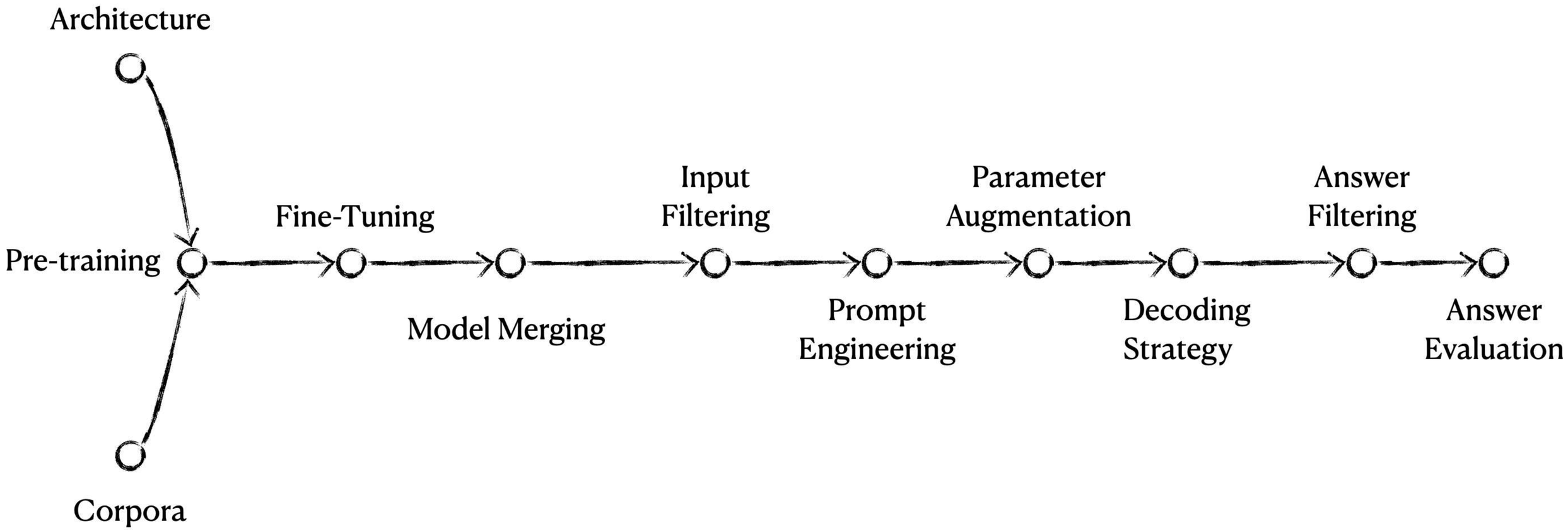
Overview of Solutions

Mapped to Model Pipeline



Overview of Solutions

Mapped to Model Pipeline



Overview of Solutions

Improving Consistency

- Fine-tuning
  - Use a different loss function
  - Train a validation head
- Knowledge plug-in
  - Dictionaries
  - New events
- Output ensemble
  - Candidate generation & candidate evaluation

Overview of Solutions

Improving Factuality

- Contrastive Decoding — seems very promising!
- ITI method for attention shifting
  - Shift attention «activations along the truth-correlated direction, which is obtained through knowledge probing».

Knowledge Conflicts

Varieties and Solutions

	Memory Coherent & Correct	Memory Coherent & Incorrect	Memory Incoherent & [In]Correct
Context Coherent & Correct	Conflict: - Answer: +	Conflict: CM Answer: ? Action: prioritize context	Conflict: IM Answer: ? Action: prioritize context
Context Coherent & Incorrect	Conflict: CM Answer: ? Action: prioritize memory	If agree: Conflict: -, answer: - If disagree: Conflict: CM, answer: -	If agree: Conflict: -, answer: - If disagree: Conflict: CM, answer: -? Action: prioritize memory + change decoding strategy
Context Incoherent & [In]Correct	Conflict: IC, Answer: -? Action: prioritize memory	If agree: Conflict: -, answer: - If disagree: Conflict: CM, answer: -? Action: filter context + prioritize context	If agree: Conflict: -, answer: ? If disagree: Conflict: CM, answer: ? Action: filter context, change decoding
No Context Provided	Conflict: - Answer: +	Conflict: - Answer: - Action: add RAG, prooritize context	Conflict: IM Answer: ? Action: change decoding

Considerations

Considerations

Overview

- Knowledge Conflicts in the wild
- Solution at a Finer Resolution
  - Work with sources of conflicts
- Evaluation on downstream tasks
  - How do Knowledge Conflicts influence NLP tasks?
- Interplay among the conflicts
- Explainability
- Multilinguality
- Multimodality

Critique

Publication

Knowledge conflicts for llms: A **survey**

R Xu, Z Qi, Z Guo, C Wang, H Wang, Y Zhang… - arXiv preprint arXiv …, 2024 - arxiv.org

… Figure 2: We view **knowledge conflict** not only as a standalone phenomenon but also as a

… Analysis, our **survey** involves systematically observing these **conflicts**, offering insights into …

☆ Save Cite Cited by 194 Related articles All 5 versions »

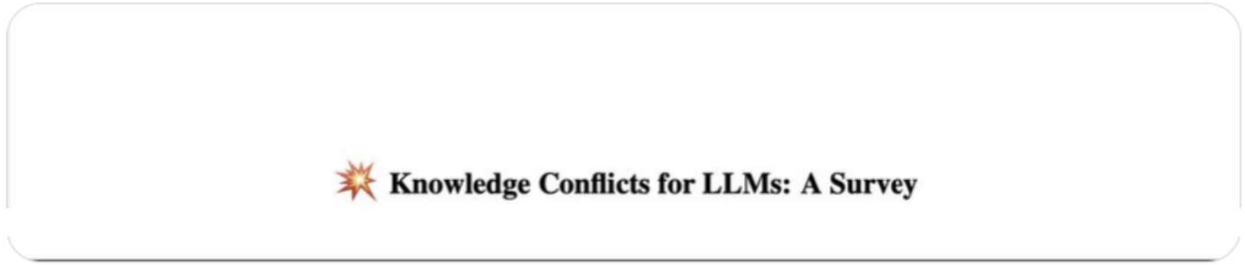
Twitter

Rongwu Xu
@rongwu_xu

🔥New paper alert☕
🧠 Navigating the Complexities of Knowledge Conflict 📚

Our latest survey delves into the heart of knowledge conflicts, revealing how LLMs tackle context-memory, inter-context, and intra-memory clashes.

📜arxiv.org/abs/2403.08319
💻github.com/pillowsofwind/...



6:07 AM · Jun 25, 2024 · **9,020** Views

GitHub Life

Knowledge-Conflicts-Survey Public

Watch 3 | Fork 7 | Star 144

main | 1 Branch | 0 Tags | Go to file t | Add file | Code

pillowsofwind Update README.md — 90ce856 · last year — 25 Commits

figures	remove DS_Store	last year
README.md	Update README.md	last year

README

💥Knowledge Conflicts for LLMs: A Survey

📢**News 9/21/2024: Our paper got accepted by EMNLP 2024 as a Main Conference Paper!**

📢**We've missed your work? Contact Rongwu directly!**

This is the repository for the survey paper: Knowledge Conflicts for LLMs: A Survey. ➡️ [中文介绍@机器之心]

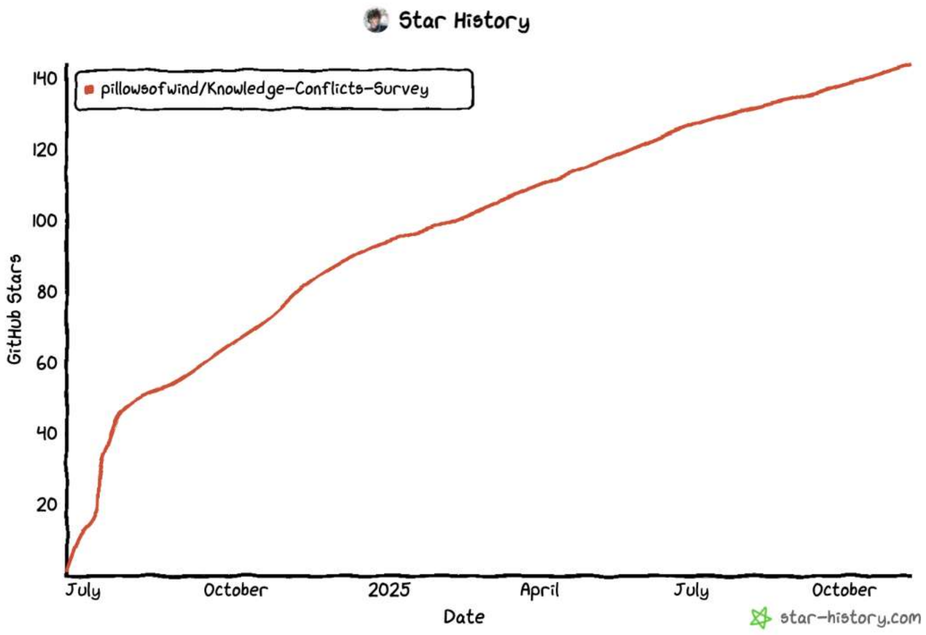
About

[EMNLP 2024] The official GitHub repo for the survey paper "Knowledge Conflicts for LLMs: A Survey"

arxiv.org/abs/2403.08319

natural-language-processing | knowledge | knowledge-representation | explainable-ai | interpretable-machine-learning | retrieval-systems | llm | chatgpt | large-language-model | retrieval-augmented-generation

Readme
Activity
144 stars
3 watching



Questions

- Can we just do positive model updates upon false info when training?
- Do you think the papers from 4 years ago can still talk about relevant problems?
- Can a knowledge conflict arise if the model is trained on perfectly coherent data?
- With what probability would the model find a proposition correct? Would it correspond to the frequency of the given fact in the training data?

References

The Paper

1. Xu, R., Qi, Z., Guo, Z., Wang, C., Wang, H., Zhang, Y. and Xu, W. (2024) 'Knowledge Conflicts for LLMs: A Survey', *arXiv preprint*, arXiv:2403.08319. Available at: https://arxiv.org/abs/2403.08319 (Accessed: 11 November 2025).

2. Xu, R. and Qi, Z. (2024) *Knowledge Conflicts for LLMs*. [Presentation slides]. Available at: https://rongwuxu.com/index_files/slides/Knowledge%20conflict.pdf (Accessed: 11 November 2025).

Knowledge and Conflicts

3. Ichikawa, J.J. and Steup, M. (2017) 'The Analysis of Knowledge', *Stanford Encyclopedia of Philosophy*. Available at: https://plato.stanford.edu/entries/knowledge-analysis/ (Accessed: 11 November 2025).

4. McLeod, E. (2011) 'Is knowledge all there is?', *Falkayn's Cogitations* [blog]. Available at: https://blog.falkayn.com/2011/03/is-knowledge-all-there-is.html (Accessed: 11 November 2025).

5. Cho, A., Kim, G.C., Karpekov, A., Helbling, A., Wang, Z.J., Lee, S., Hoover, B. and Chau, D.H. (2024) 'Transformer Explainer: Interactive learning of text-generative models', *arXiv preprint*, arXiv:2408.04619. Available at: https://arxiv.org/pdf/2408.04619.pdf (Accessed: 11 November 2025).

6. Marr, B. (2020) 'How artificial intelligence is changing science', *Nature*, 579(7797), pp. 24–26. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC7179339/ (Accessed: 11 November 2025).

References

Knowledge and Graphs

7. Karpekov, A., Kim, G.C., Hoover, B., Xu, J., Nashaat, M., Kahng, M., Chau, D.H. and Xu, C. (2024) 'Learning to simulate: A data-centric generative agent framework for digital twins', in *2024 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Available at: **https://ieeexplore.ieee.org/abstract/document/11021395** (Accessed: 11 November 2025).

8. Shi, X., Li, Q., Chen, L., Yan, Z., Zhao, W. and Wang, S. (2019) 'A novel time constraint-based approach for knowledge graph conflict resolution', *International Journal of Computational Intelligence Systems*, 12(2), pp. 1112–1126. Available at: **https://www.researchgate.net/publication/336635969_A_Novel_Time_Constraint-Based_Approach_for_Knowledge_Graph_Conflict_Resolution** (Accessed: 11 November 2025).

9. Liu, X., Gao, Y., Xu, R., Chen, Q., Li, L. and Wang, Q. (2024) 'A knowledge conflict detection method for large-scale knowledge graphs', *Procedia Computer Science*, 242, pp. 1757–1765. Available at: **https://www.sciencedirect.com/science/article/pii/S1877050924030072** (Accessed: 11 November 2025).

10. Qi, N., Zeng, Z., Li, P. and Ma, Z. (2024) 'Applications of knowledge graph and dependent interpretations in mathematical question answering', *Mathematics*, 12(15), p. 2318. Available at: **https://www.mdpi.com/2227-7390/12/15/2318** (Accessed: 11 November 2025).

11. Qian, C., Zhao, X. and Wu, S.T. (2023) '"Merge Conflicts!" Exploring the impacts of external distractors to parametric knowledge graphs', *arXiv preprint*, arXiv:2309.08594. Available at: **https://arxiv.org/pdf/2309.08594** (Accessed: 11 November 2025).

References

Model Behaviours

12. Elazar, Y., Kassner, N., Ravfogel, S., Ravichander, A., Sade, A., & Goldberg, Y. (2021) 'Measuring and Improving Consistency in Pretrained Language Models', *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8069–8087. Available at: https://aclanthology.org/2021.emnlp-main.636/ (Accessed: 11 November 2025).

13. Li, H., Reif, E., Bai, Y., Ma, S., Liu, C., Yang, Z., Zheng, H., Song, D., Liang, P., Schuurmans, D., Wang, X., and Tan, X. (2023) 'Benchmarking and improving generator-validator consistency of LMs', *OpenReview*. Available at: https://openreview.net/pdf?id=phBS6YpTzC (Accessed: 11 November 2025).

14. Tenney, I., Das, D. and Pavlick, E. (2019) 'BERT rediscovers the classical NLP pipeline', *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601. Available at: https://aclanthology.org/P19-1452.pdf (Accessed: 11 November 2025).

15. Xue, B., Wang, H., Wang, R., Wang, S., Wang, Z., Du, Y., Liang, B. and Wong, K.-F. (2024) 'A comprehensive study of multilingual confidence estimation on large language models', *arXiv preprint*, arXiv:2402.13606. Available at: https://arxiv.org/abs/2402.13606 (Accessed: 11 November 2025).

Inter-Context Conflicts

16. Li, J., Wang, Q., Saleh, M. and Jia, R. (2024) 'ContraDoc: Understanding self-contradictions in documents with large language models', Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 6509–6523. Available at: https://aclanthology.org/2024.naacl-long.362.pdf (Accessed: 11 November 2025).

17. Zhang, M.J.Q. and Choi, E. (2021) 'SITUATEDQA: Incorporating extra-linguistic contexts into QA', *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7371–7387. Available at: https://aclanthology.org/2021.emnlp-main.586.pdf (Accessed: 11 November 2025).

References

Inter-Context Conflicts

...

18. Kasai, J., Sakaguchi, K., Takahashi, Y., Le Bras, R., Asai, A., Yu, X.V., Radev, D., Smith, N.A., Choi, Y. and Inui, K. (2024) 'REALTIME QA: What's the answer right now?', *arXiv preprint*, arXiv:2207.13332. Available at: https://arxiv.org/pdf/2207.13332 (Accessed: 11 November 2025).

19. Li, J., Raheja, V. and Kumar, D. (2024) 'CONTRADOC: Understanding self-contradictions in documents with large language models', *arXiv preprint*, arXiv:2311.09182. Available at: https://arxiv.org/pdf/2311.09182 (Accessed: 11 November 2025).

20. Li, J., Raheja, V. and Kumar, D. (2024) 'ContraDoc: Understanding self-contradictions in documents with large language models', Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 6509–6523. Available at: https://aclanthology.org/2024.naacl-long.362.pdf (Accessed: 11 November 2025).

Intra-Memory Conflicts

21. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B. and Liu, T. (2024) 'A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions', *arXiv preprint*, arXiv:2311.05232. Available at: https://arxiv.org/pdf/2311.05232 (Accessed: 11 November 2025).

22. Heidloff, N. (2023) 'Decoding methods for generative AI', *Niklas Heidloff – Blog*. Available at: https://heidloff.net/article/greedy-beam-sampling/ (Accessed: 11 November 2025).

References

Intra-Memory Conflicts

...

23. Wong, Y.C., Rizvi, S.T.H., Lau, T., Ahmad, N., Bouktif, S. and Guan, C. (2023) 'Determining the best feature combination through text and probabilistic feature analysis for GPT-2-based mobile app review detection', *Expert Systems with Applications*, 238, 122418. Available at: **https://www.researchgate.net/publication/376832794_Determining_the_best_feature_combination_through_text_and_probabilistic_feature_analysis_for_GPT-2-based_mobile_app_review_detection** (Accessed: 11 November 2025).

24. Yao, Y., Wang, P., Tian, B., Cheng, S., Li, Z., Deng, S., Chen, H. and Zhang, N. (2023) 'Editing large language models: Problems, methods, and opportunities', *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10222–10240. Available at: **https://aclanthology.org/2023.emnlp-main.632/** (Accessed: 11 November 2025).

Context-Memory Conflicts

25. Fatemi, B.B. (2025) 'LLMs: the object or the lens', *1st Workshop on Large Language Models for Cross-Temporal Research (XTempLLMs 2025)*, YouTube, 10 October. Available at: https://www.youtube.com/watch?si=qZhQ5GW-YI9XvDEg&t=11624&v=Um27xolvBfc&feature=youtu.be (Accessed: 11 November 2025).

26. Wang, Y., Feng, S., Wang, H., Shi, W., Balachandran, V., He, T. and Tsvetkov, Y. (2024) 'Resolving knowledge conflicts in large language models', *Proceedings of the 1st Conference on Large Language Models (COLM 2024)*. Available at: **https://arxiv.org/pdf/2310.00935** (Accessed: 11 November 2025).

27. Chuang, Y.-S., Xie, Y., Luo, H., Kim, Y., Glass, J. and He, P. (2024) 'DoLa: Decoding by contrasting layers improves factuality in large language models', *Proceedings of the International Conference on Learning Representations (ICLR 2024)*. Available at: **https://arxiv.org/pdf/2309.03883** (Accessed: 11 November 2025).