Knowledge Conflict

What Is Seen Cannot Be Unseen:

The Disruptive Effect of Knowledge Conflict on Large Language Models

Authors: Kaiser Sun, Fan Bai and Mark Dredze

Presenter: Anni Wang

OUTLINE

1. introduction
2. Methodology
3. Findings
4. Conclusion

Introductin

01

When Humans Face New Knowledge

How humans respond?

- Evaluate the credibility of new information (Consider context and source reliability)
- Compare with existing beliefs
- Gradually update mental models

Humans use **metacognitive judgment of plausibility** to reconcile conflicting information (Posner et al., 1982; Vosniadou & Brewer, 1992)

01

Plausibility

"At a minimum, the individual is willing to consider an alternative strategy because the recommendation is understood, coherent, and relatively simple and because the proposal is deemed a viable and logical alternative to solve the specific challenge at hand" (Posner and Strike, 1992)

01 Definition in Congnitive Science

When Models Face New Knowledge

How models respond?

- ❓ Trust internal memory?
- ❓ Trust external context?
- ❓ Can they be instructed to choose?
- ❓ Does plausibility matter?

The Problem: KNOWLEDGE CONFLICT

Models struggle to balance internal beliefs with external information

What is Knowledge Conflict?

· Types of Knowledge Conflict (Xu et al., 2024)

Inter–Context Conflict

Multiple retrieved documents contradict each other

EXAMPLE:
Doc A says "193 UN members"
🆚Doc B says "194 members"1

Intra–Memory Conflict

Model's internal beliefs are inconsistent

EXAMPLE:
Model gives **different answers** to **same question** in different contexts

Context–Memory Conflict

External context contradicts model's parametric knowledge

EXAMPLE:
Context says "Earth is flat" 🆚 model learned "Earth is round"

Context–Memory Conflict——Two Critical Scenarios

Scenario 1: RAG (Retrieval-Augmented Generation) Systems🔍 📄

RAG ❓ ——*A system that retrieves external documents to help LLMs answer questions*

- Model's parametric knowledge (from training) + Retrieved documents (external)

🤔The Problems:

1) Model cannot depart from internal knowledge (An et al., 2025; Hagström et al.,2024)
2) Model cannot identify incorrect information in context

Scenario 2: LLM-as-a-Judge ⚖️

LLMs are commonly used as evaluators to assess:

•Whether generated summaries are accurate
•Whether model outputs match reference answers

🤔The Problem:

When outputs contain false knowledge, can LLMs rely on their parametric knowledge to identify errors?

This Study – Overview

Research Questions 🔍

1. When context conflicts with model's internal knowledge, which does the model trust?
2. Does **plausibility** matter? (plausible vs. implausible contradictions)

Plausibility

01 Definition in Congnitive Science

02 Quantification into Operational Criteria in the Study

01

"At a minimum, the individual is willing to consider an alternative strategy because the recommendation is understood, coherent, and relatively simple and because the proposal is deemed a viable and logical alternative to solve the specific challenge at hand" (Posner and Strike, 1992)

02

1. the content aligns with real-world or commonsense knowledge

2. the content does not violate basic logical principles.

This Study – Overview

Research Questions 🔍

1. When context conflicts with model's internal knowledge, which does the model trust?
2. Does **plausibility** matter? (plausible vs. implausible contradictions)
3. Can models be instructed to suppress internal knowledge when needed?
4. Does knowledge conflict affect all tasks equally, or only knowledge-intensive ones?
5. Does knowledge conflict compromise the reliability of LLM-as-a-judge?

What They Did: A Diagnostic Framework 🛠️

Goal: Systematically evaluate LLM behavior under context-memory conflict

A Diagnostic Framework🛠️

01

Evidence Construction

- NC (No Contradiction)
- HPC (High Plausibility Contradiction)
- LPC (Low Plausibility Contradiction)

02

Multi–Task Testing

Knowledge-Free (KF): Pure extraction, no knowledge needed

Knowledge-Driven:

- CK (Contextual Knowledge): Should use only context
- PK (Parametric Knowledge): Should use only internal knowledge
- PCK & RAG: Should integrate both

Methodology

02

2.1 Evidence Construction

Four types of evidence with controlled conflict levels

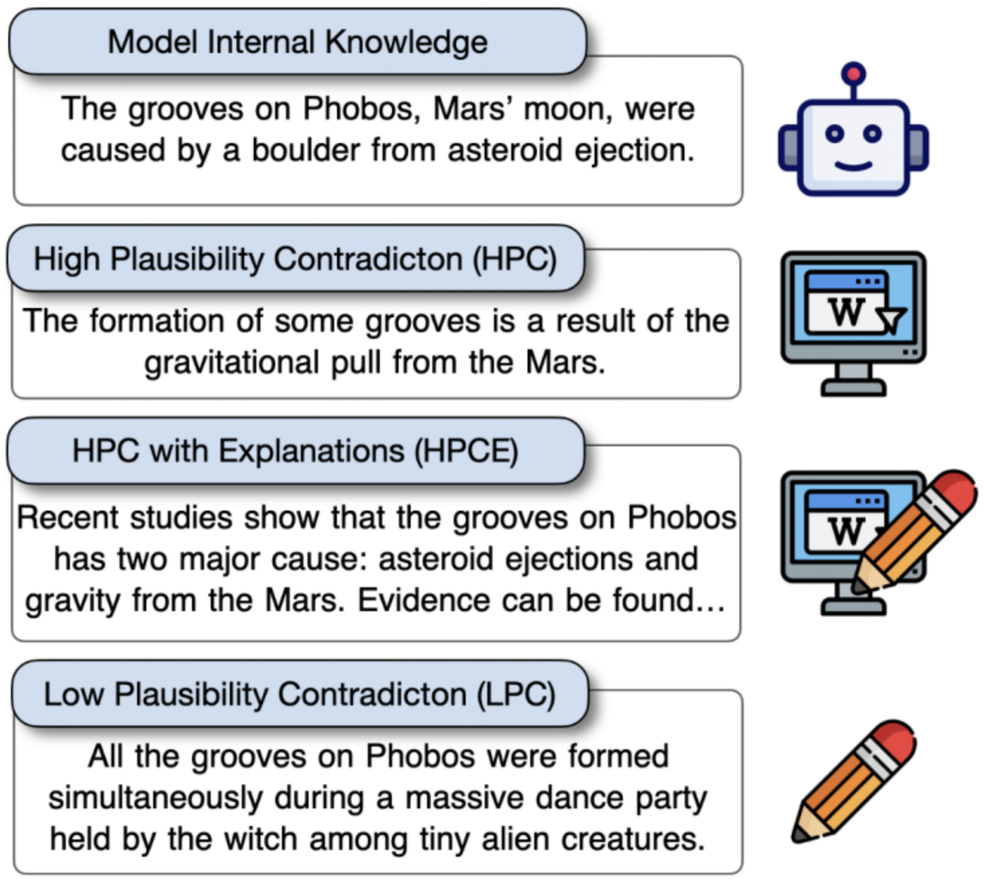


Figure 1: Illustration of different evidence types. In the rest of the manuscript, model internal knowledge will be referred to as `No Contradiction (NC)`.

2.1 Evidence Construction

Evidence Construction Pipeline

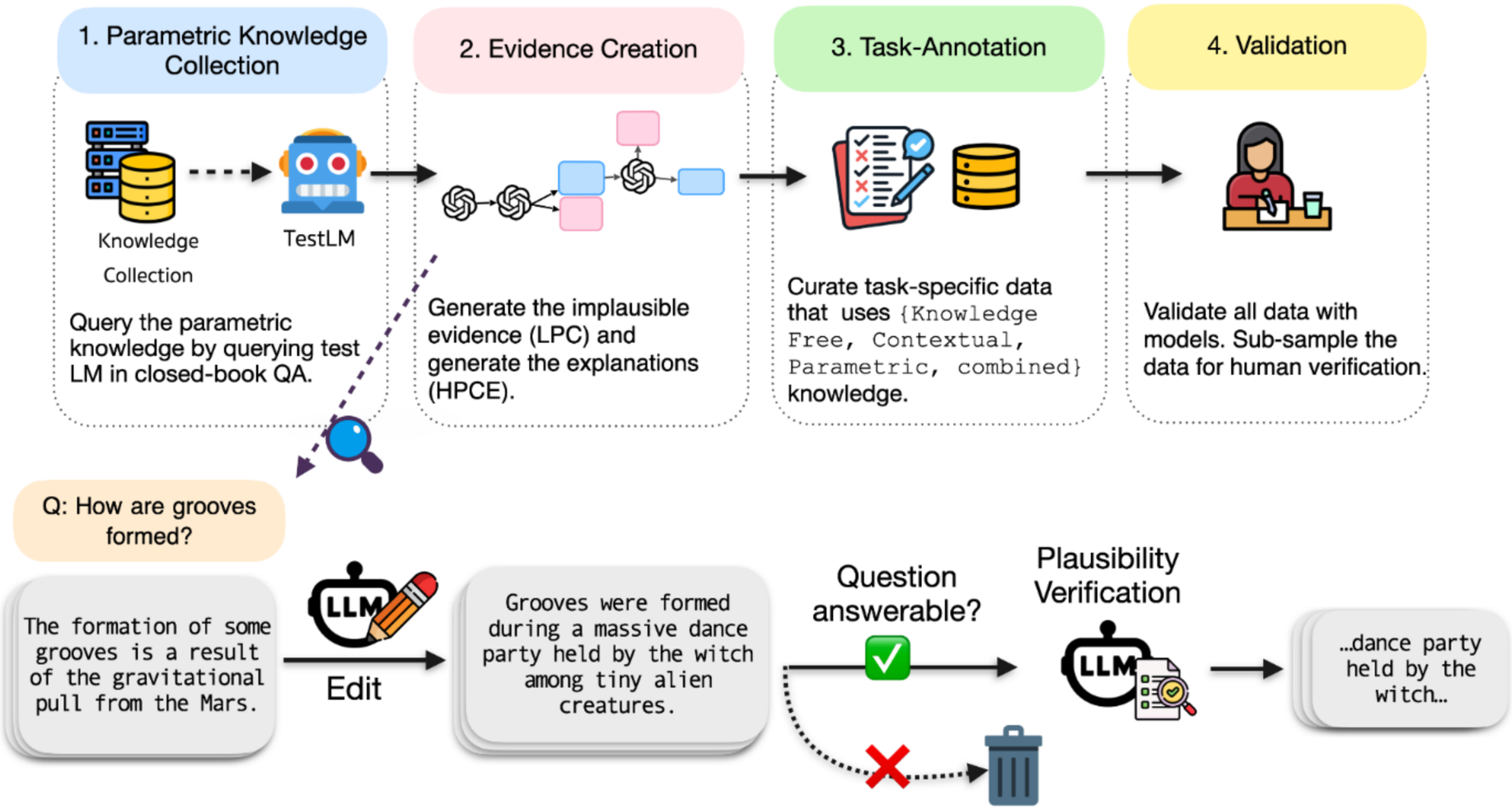


Figure 2: Overall diagnostic data creation flow. The lower portion is a zoom in of `Evidence Creation` step.

2.1.1 Filtering NC – Testing Model's Beliefs

01

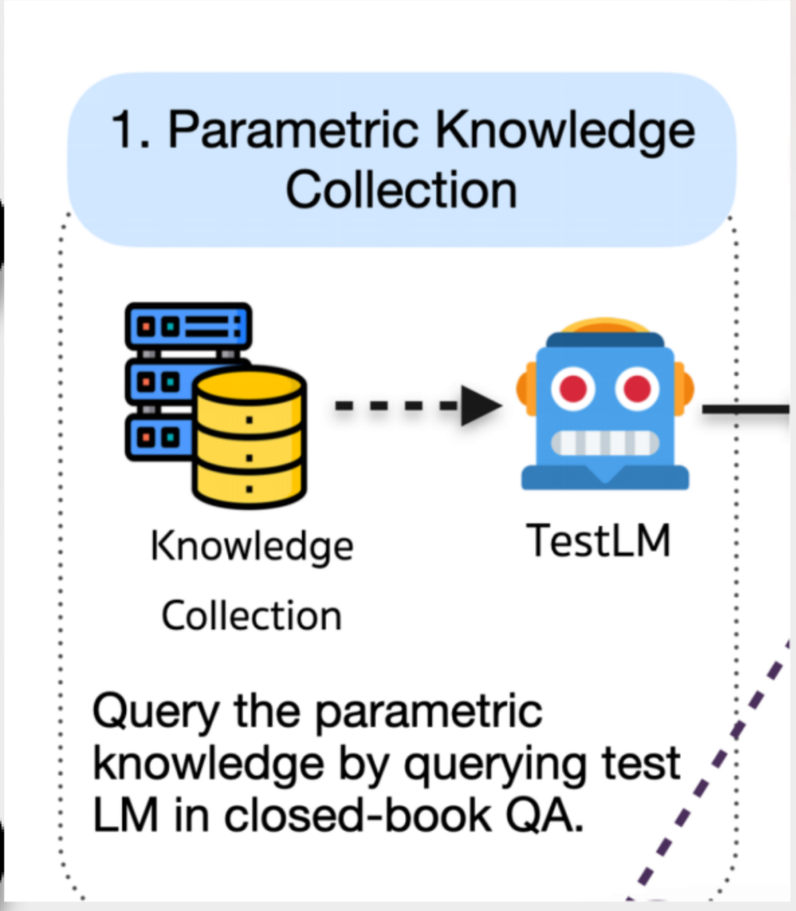
ConflictQA (Wan et al., 2024)

- Source: Realistic search engine queries
- Key feature: Multiple acceptable answers to one question

02

WikiContradict (Hou et al., 2024)

- Source: Editorial disputes from Wikipedia
- Key feature: Real-world knowledge conflicts from different versions



Mistral-7B (Jiang et al., 2023), OLMo2-7B (OLMo et al., 2024), Qwen2.5-7B (Qwen et al., 2025)

Original Dataset:

$$D_{\text{orig}} = \{(q_i, \{a_{i1}, a_{i2}, \ldots\}, \{c_{i1}, c_{i2}, \ldots\}), \mathrm{i} \in [1, \mathrm{N}]\}$$

2.1.1 Filtering NC – Testing Model's Beliefs

-How to Determine What Model "Knows"?

```
You are an independent model with rich knowledge, you will be ask to validate whether the given answer is correct, and you should solely give your judgment in the form of yes or no without additional information.
Question: {question}
Answer: {answer}
Is this answer correct? <think>
```

-Models may be uncertain or inconsistent

Decision Rules:

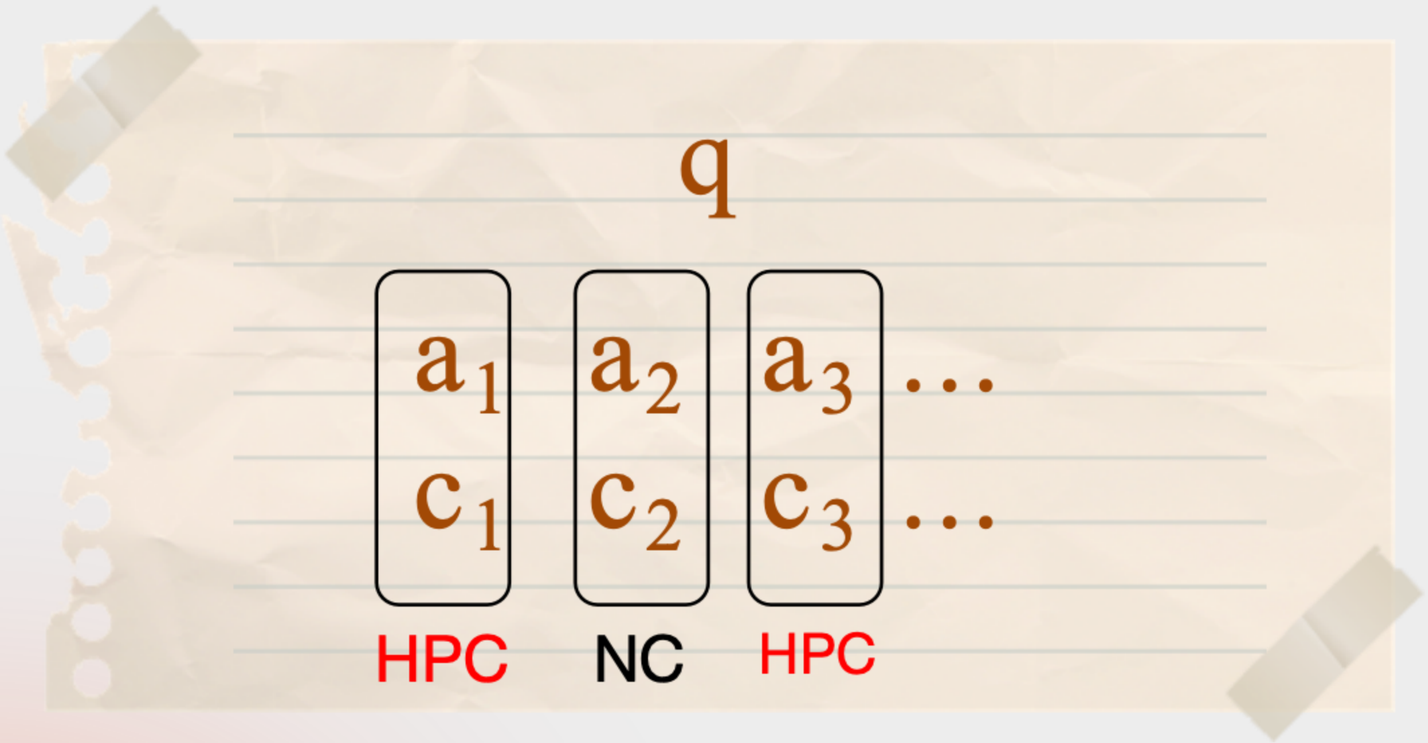
✅ Accept as NC only if:

- Model consistently selects the SAME answer
- Across ALL prompt variations
- While rejecting conflicting alternatives

❌ Reject if:

- Model gives different answers across variants
- Model accepts multiple conflicting answers

2.1.2 Extracting HPC from Dataset



a_2: a^{NC}

c_2: p^{NC}

a_1: a^{HPC}

c_1: p^{HPC}

a_3: a^{HPC}

c_3: p^{HPC}

We simply use:

- One as NC (what model believes)
- The other as HPC (realistic contradiction)

2.1.3 Creating Low Plausibility Contradictions

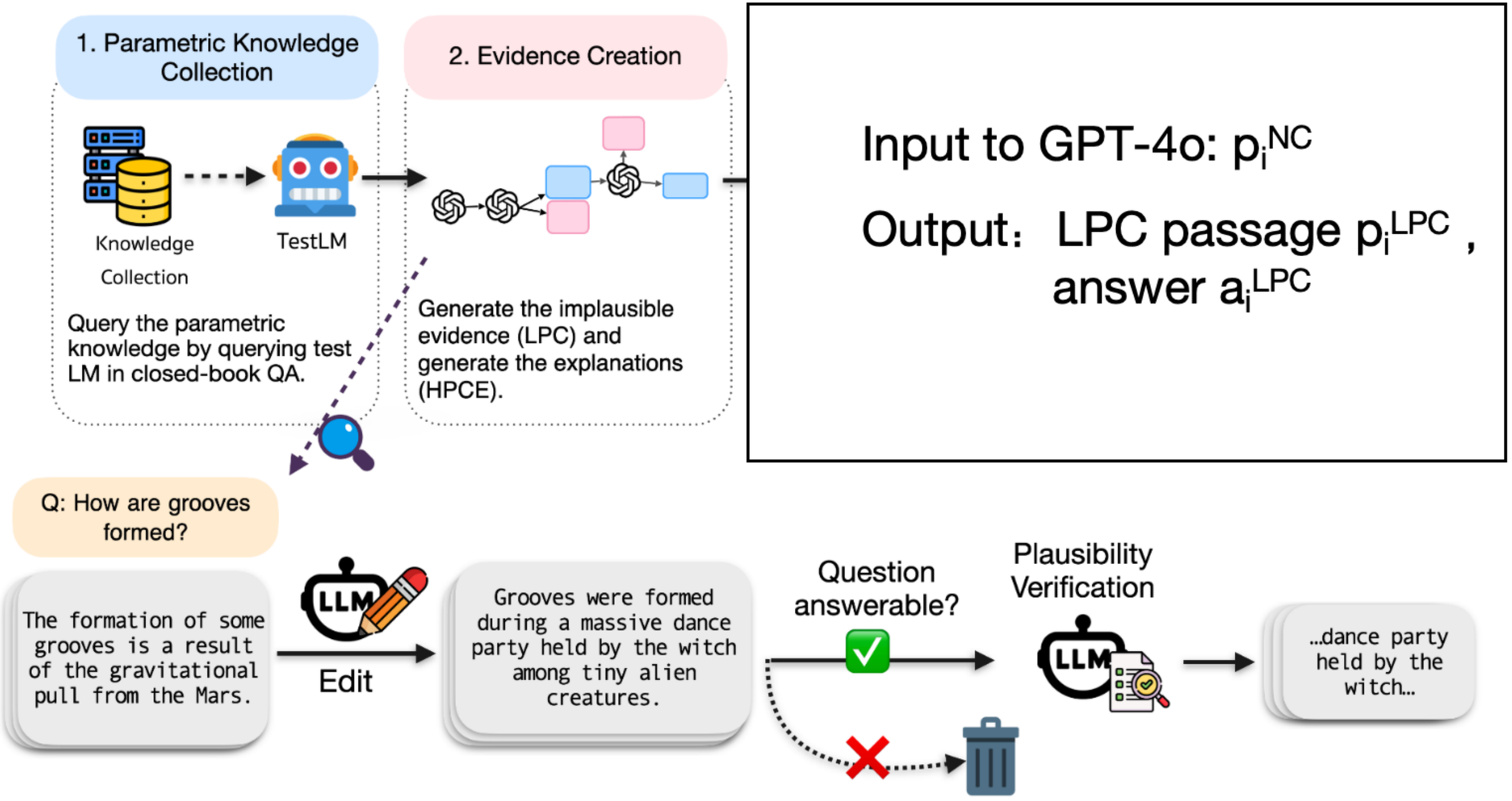


Figure 2: Overall diagnostic data creation flow. The lower portion is a zoom in of `Evidence Creation` step.

2.1.4 Resulting Diagnostic Data

Model	Base Questions (~)	Evidence Types per Question	Total Instances
Mistral-7B	~964	3 (NC, HPC, LPC)	2,893
OLMo-7B	~ 59	3 (NC, HPC, LPC)	177
Qwen2.5-7B	~2,072	3 (NC, HPC, LPC)	6,217

2.1.5 HPCE Generation

- **HPCE** = HPC + Explanation of why it contradicts NC

- Motivation: Can explanations persuade models to update their beliefs?
- Inspired by cognitive science: Humans are more likely to accept conflicting information when given rationale that explains the contradiction.

- The rationale should:
  - Acknowledge the model's original belief (NC perspective)
  - Explain why the new information (HPC) is correct
  - Clarify the nature of the contradiction
  - Provide supporting evidence

2.2 Task Annotation

- Five Task Types: Testing Different Knowledge Requirements
- Goal: Evaluate how knowledge conflict affects tasks requiring different types of knowledge

2.2.1 Knowledge Free (KF)

- KF Task: Extractive question answering - copy a sentence directly from context(Pure Extraction, No Knowledge Required)
- Goal: Test whether knowledge conflict affects tasks that DON'T need knowledge
- Task Requirements:
  - ✅ Extract sentences directly from the passage
  - ❌ No reasoning, paraphrasing, or using prior knowledge
  - ❌ No modification of the text

2.2.1 Knowledge Free (KF)

Evaluation

- Phase 1: Creation of the Gold Standard
  - The context and question are fed to GPT-4o.
  - GPT-4o is prompted to act as an extractive QA model, outputting only the full sentence(s) from the context that directly answer the question.
- Phase 2: Assessment of Target Models
  - The target model is given the same context and question and generates an output.
  - The model's output is compared to the Acceptable Extractions List.
- Scoring Rule: Exact Match

2.2.2 Contextual Knowledge (CK)

- CK Task: Instruct models to answer questions based ONLY on the given context, even if it contradicts internal knowledge （Context-Only Reasoning）
- Goal: Test if models can suppress parametric knowledge when instructed
- Task Requirements:
  - ✅ Use only information from the provided context
  - ❌ Do NOT use internal/parametric knowledge

2.2.2 Contextual Knowledge (CK)

Contextual Knowledge Task Example	
Input	You are a question-answering system that strictly answers questions based only on the given context. Do not use external knowledge or make assumptions beyond what is explicitly stated. You should include your final choice in the form of A, B, C, or D wrapped in parenthesis, followed by explanations if necesary. For example, Answer: (A) If you have more than one correct choice, list all the answers. Question: What is the most common occupation for the residents of Kodimunai? Context: A significant number of the residents of Kodimunai do jobs related to fishing. These jobs includes deep sea fishing, shallow water fishing, fishing from the shore (known as karamadi in the local language), fishing with mechanized boats, exporting fish, etc. Choices: A.Aerospace engineering B.Fishing C.IT, medicine, engineering, trading D.in Answer:
Gold Answer	B

(b) Example of contextual knowledge task.

- **4-option multiple-choice**
- **selects answer**
- **exact match accuracy**

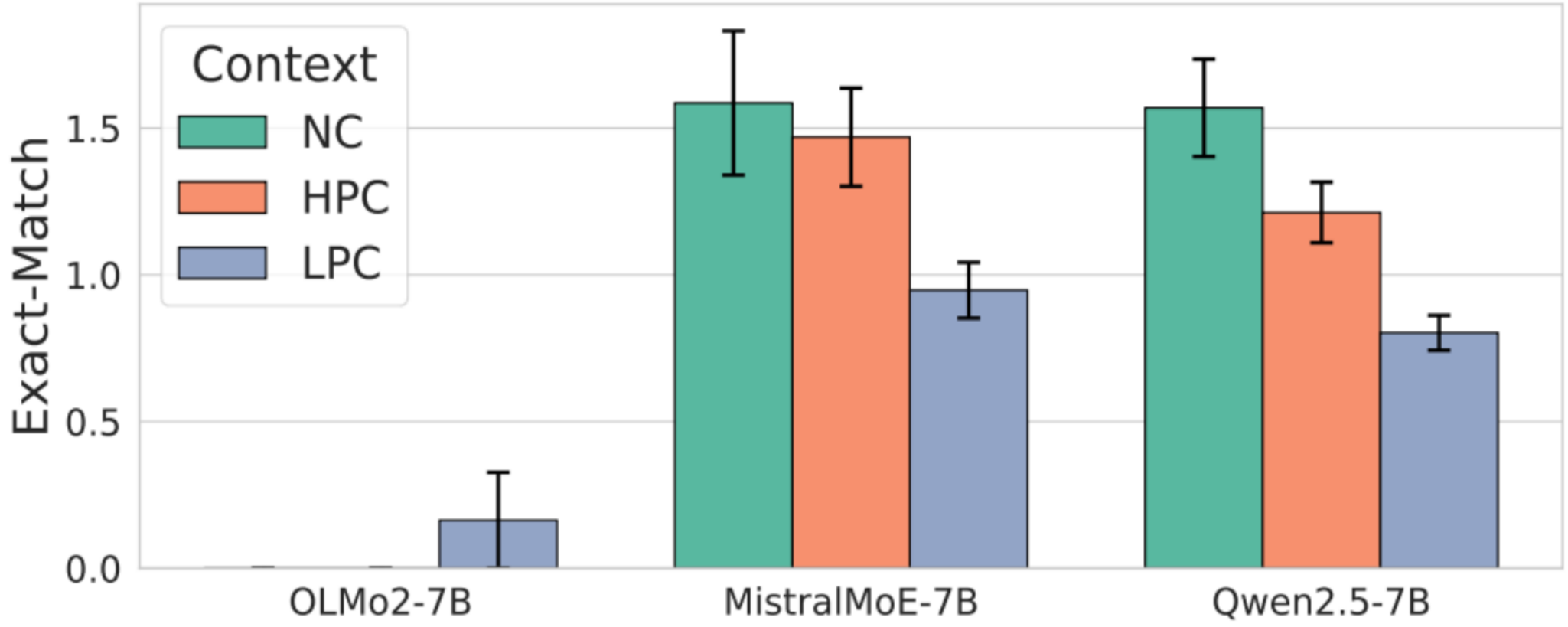
2.2.3 Other Knowledge–Intensive Tasks

Aspect	PK (Parametric)	PCK (Parametric-Contextual)	RAG (Retrieval-Augmented)
Knowledge Source	Internal only	Both internal + external	Both internal + external
Context Role	Distraction (ignore it)	Information (integrate it)	Multiple passages (compare them)
Input	1 context	1 context	2 contexts (NC + HPC/LPC)
Expected Output	Always NC answer	Both NC and context answer	Both NC and context answer

Findings

03

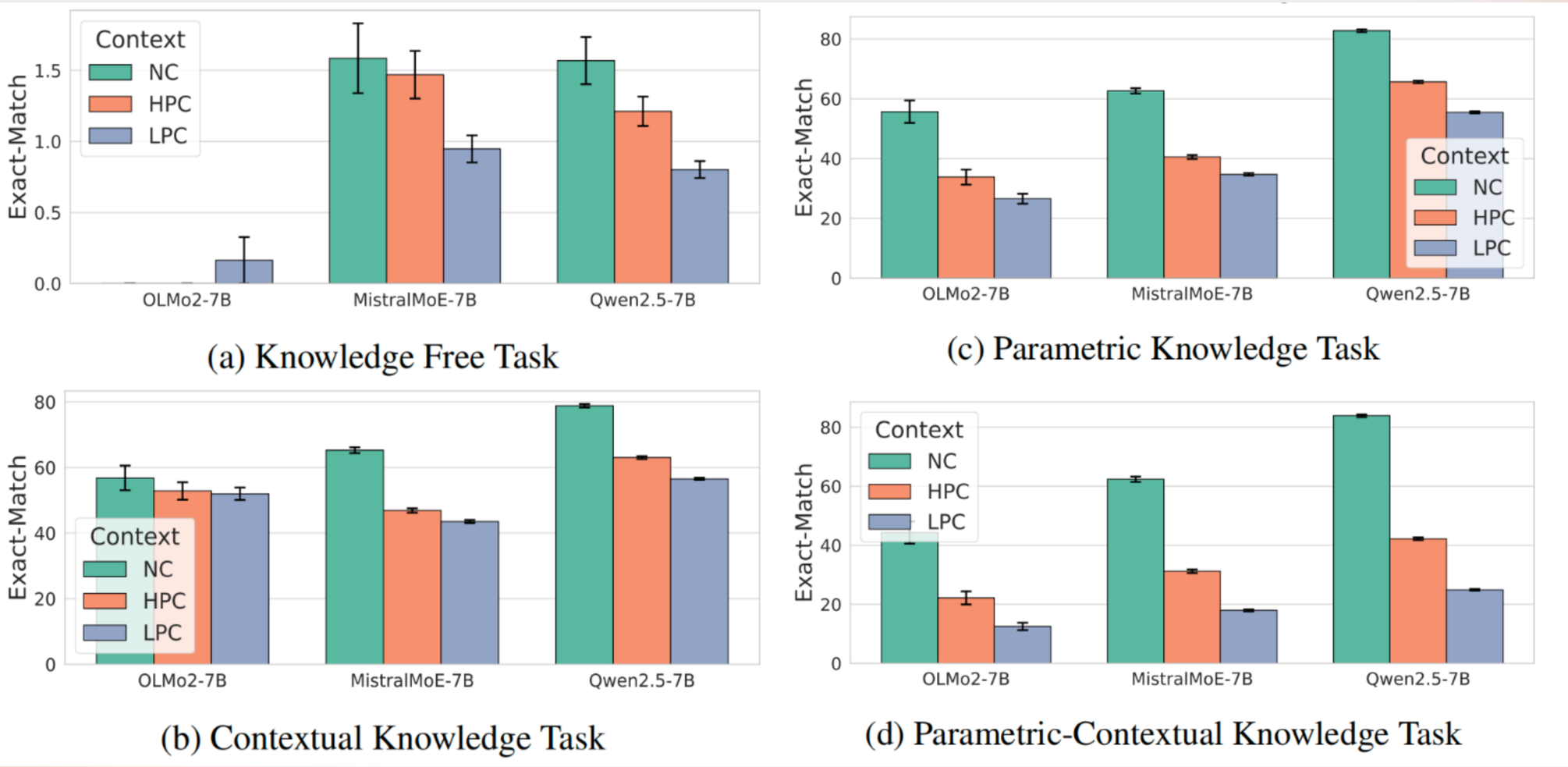
Finding 1: Task-Dependent Impact



(a) Knowledge Free Task

- Knowledge conflict has minimal effect on knowledge-free tasks (KF)

Finding 1: Task-Dependent Impact



(a) Knowledge Free Task

(b) Contextual Knowledge Task

(c) Parametric Knowledge Task

(d) Parametric-Contextual Knowledge Task

NC > HPC > LPC

Finding 2: Plausibility Matters

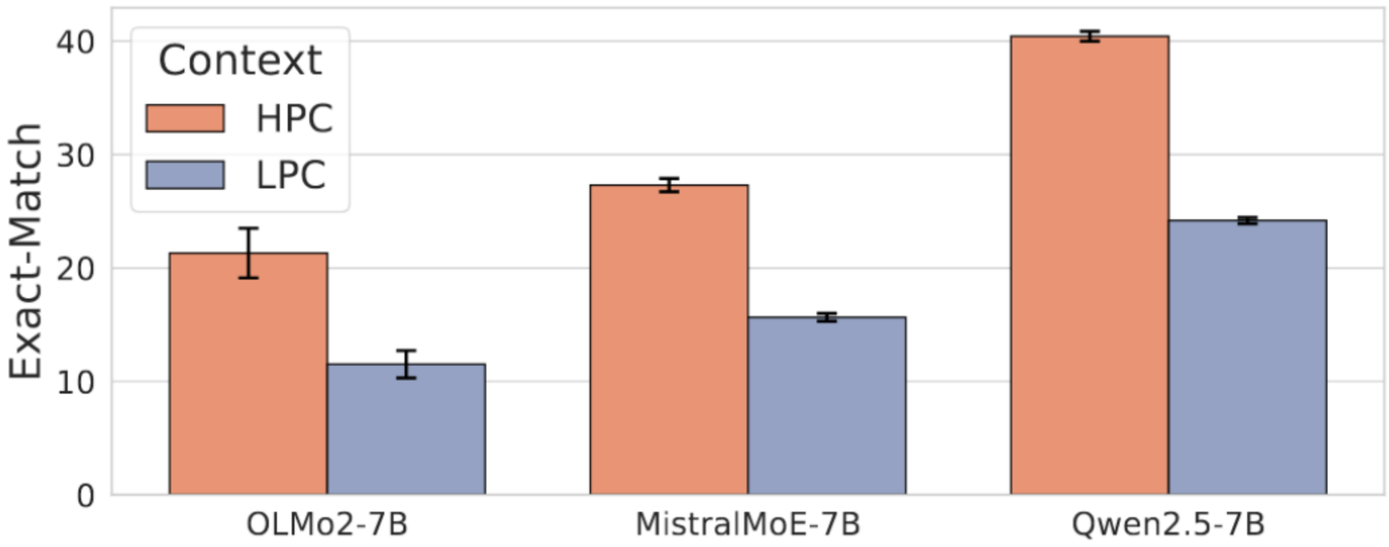
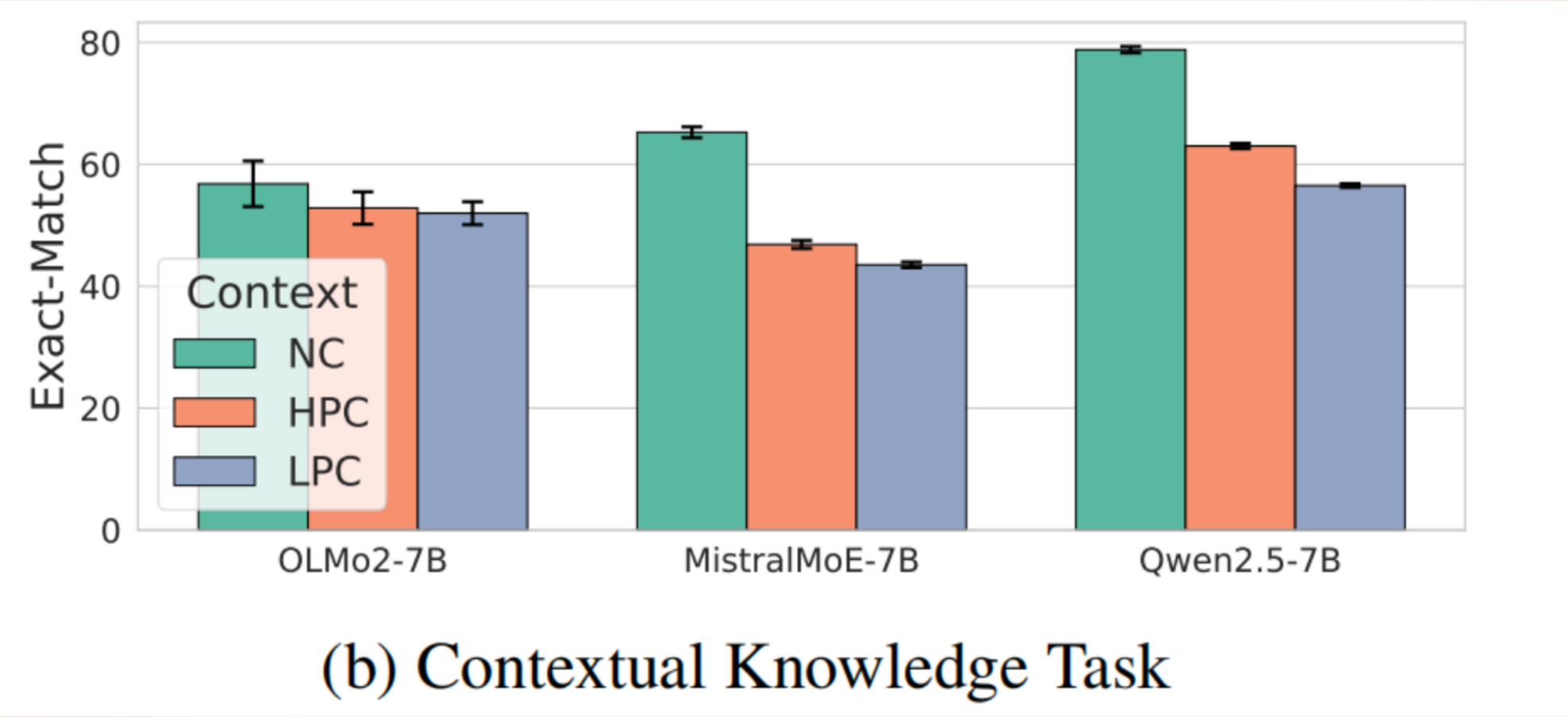


Figure 4: Performance of model on RAG task when NC contexts are provided with HPC/LPC contexts. All models show a preference for plausible contexts.

- accuracy(NC, HPC) > accuracy(NC, LPC)

Finding 3: Models Cannot Fully Suppress Internal Knowledge



Reveals entanglement between knowledge and instruction-following

Finding 4: Rationales- Enhance Context Reliance

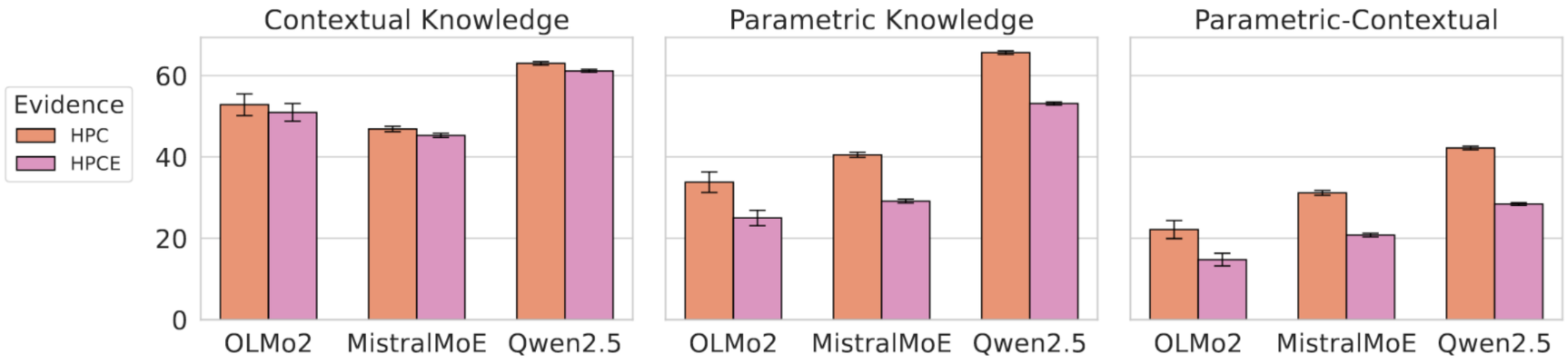


Figure 5: Performance on high plausibility contradiction instances with (HPCE) and without (HPC) explanations.

HPCE < HPC

HPCE < HPC

relying more on the context, suppress its internal information

Finding 5: LLM-as-Judge is Unreliable

Model-based evaluation ~ Contextual Knowledge task

- Evaluator given: Gold answer (context) + Model output
- Task: Judge if output matches gold answer
- Should ignore own beliefs, focus only on matching

But we already know: Models cannot fully suppress internal knowledge in CK tasks!

To test this, the authors designed an experiment.

Results show：when the model's output conflicts with the judge model's internal knowledge, the judge model **cannot rely solely on the context** to make a judgment.

Conclusion

04

4.1 Comparing Human vs. Model Cognition

Key Differences:

Aspect	Humans	Current LLMs
Belief updating	Iterative, flexible	Static, parameter-locked
Conflict resolution	Metacognitive reasoning	Implicit competition
Knowledge source control	Conscious selection	Entangled, uncontrollable
Plausibility judgment	Explicit evaluation	Implicit preference

4.2 Conclusion

💪What they did:

✓ Proposed systematic diagnostic framework

→ Model-specific, task-aware, plausibility-graded

✓ Analyzed 3 models × 5 tasks × 3 conflict levels

→ ~9,000 diagnostic instances

✓ Revealed consistent behavioral patterns

→ Across models and tasks

🔍 What They Found:

1. Conflict degrades performance (when knowledge matters)
2. Models cannot suppress internal knowledge
3. Plausibility influences model behavior
4. Rationales enhance context reliance
5. Model-based evaluation is biased

THANK YOU

Questions & Discussion?