Do LLMs actually use relations in a KG?

Evidence from Entity and Relation Perturbation

Dominic Okonkwo · Ismailcem Budak Arpinar

NeSLab · School of Computing · University of Georgia

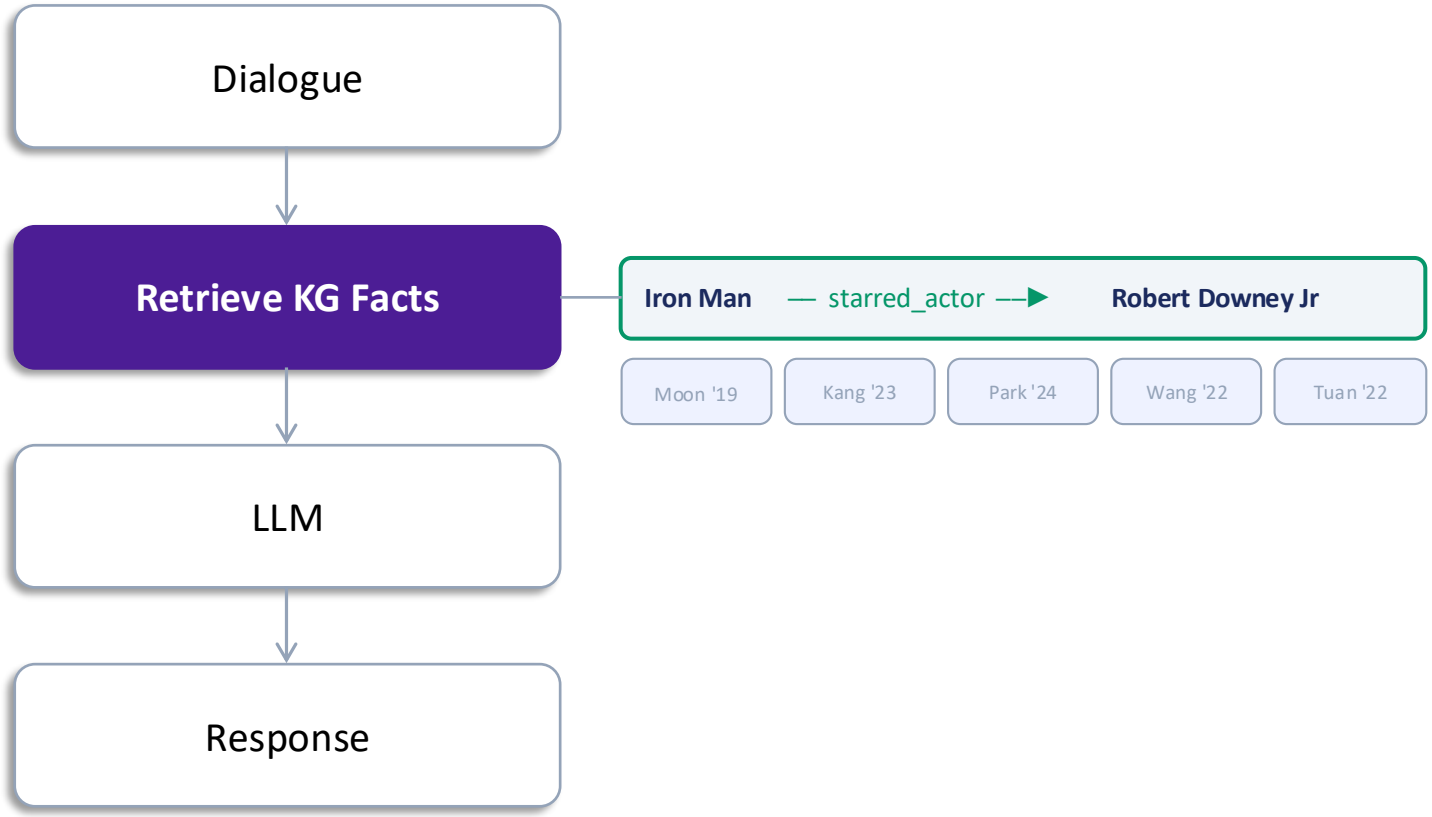
2026 SIGdial Atlanta, GA, USA

Knowledge Graphs are becoming a key component of LLM systems.

KG-grounded dialogue · KG-LLM integration · Graph reasoning, well-cited, actively growing research area

THE UNTESTED ASSUMPTION

KG-dialogue systems assume the LLM understands and uses the relation.



The gap: All of these systems build relation types into the model, but none test whether the model actually uses the typed label at inference.

SO WE ASKED

Do models actually use the relation, or just the entities?

These systems produced good results.

Researchers assumed KG grounding helped.

> But when a model generates a response, you cannot tell from the output alone whether it used the relation label, or whether it already knew the answer from training.

If the relation drove the response
removing or corrupting it should change the output

If the entity dominated
the output looks the same regardless of the relation

Nobody had directly tested which one it is.

Original KG triple

```
( Iron Man, starred_actor,
Robert Downey Jr. )
```

LLM generates

"Robert Downey Jr. starred in Iron Man, the 2008 Marvel superhero film…"

❓ **Did it use starred_actor?**

Or did it already know Robert Downey Jr. starred in Iron Man from pretraining ,making the relation label irrelevant?

METHOD — THREE CONDITIONS

Our idea: hold everything constant. Change only the relation.

Any change in response is caused by the relation, nothing else.

C1 Intact

Iron Man

starred_actors

Robert Downey Jr.

Correct relation · Full KG triple

C2 Shuffled

Iron Man

written_by

Robert Downey Jr.

Same entities · Corrupted relation

C3 Entities Only

Iron Man

—

Robert Downey Jr.

No relation · Entity baseline

If model uses the relation → C1 and C2 responses diverge.

If model ignores the relation → C1 and C2 responses look identical.

C3 = ceiling: what "no relation" looks like — maximum disruption.

Evaluation on Two Knowledge Graphs

OpenDialKG

Widely used benchmark

Human-human dialogues grounded in Freebase paths.

Freebase · 175 relations · 3,216 test turns

Relation style: short, close to natural language

`starred_actors` · `written_by` · `has_genre`

4 domains: movies, books, sports, music

WoW-KG (ours)

Built from Wikidata, independent testbed

Wizard of Wikipedia augmented with Wikidata triples.

Wikidata · 219 relations · 778 test turns

Tests whether findings generalise beyond one benchmark

`instance_of` · `occupation` · `country`

More abstract labels, harder to guess from wording alone

6 Models · 3 Families · Same Zero-Shot Prompt

Large and small variants of each family, to test whether scale changes the finding.

OpenAI

- **GPT-4o**
- GPT-4o-mini

Meta

- **Llama-3.1-70B**
- Llama-3.1-8B

Alibaba

- **Qwen2.5-32B**
- Qwen2.5-7B

Zero-shot. No fine-tuning. No few-shot examples. Same prompt template across every model and every condition.

The Prompt

Identical across all 6 models and all 3 conditions. Only the {knowledge} field changes.

SYSTEM

```
You are a knowledgeable conversational assistant.
Use the knowledge to continue the conversation naturally.
Respond in 1–3 sentences only.
```

USER

```
Topic: {dialogue topic}
Knowledge:
```

C1 `Iron Man --[starred_actor]--> Robert Downey Jr.`

C2 `Iron Man --[written_by]--> Robert Downey Jr.`

C3 `Entities: Iron Man, Robert Downey Jr.`

```
Continue the conversation as the Assistant:
```

How do we measure relation sensitivity?

If changing only the relation barely changes the response, the model likely did not use the relation.

Relation Sensitivity Score (RSS)

```
RSS = 1 − ROUGE-L( ResponseOriginal , ResponsePerturbed )
```

RSS ≈ 0 Responses nearly identical
→ **Low relation sensitivity**

RSS ≈ 1 Responses very different
→ **High relation sensitivity**

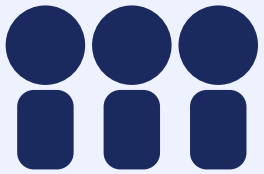
Also computed:

Div-C3 Divergence when relation is removed entirely (C3 baseline)

Ratio = RSS ÷ Div-C3 → near 1.0 means shuffling ≈ removing

We also asked Humans

RSS is computed automatically, but we also validated it with blinded human judgments.



3 independent raters

100

turns

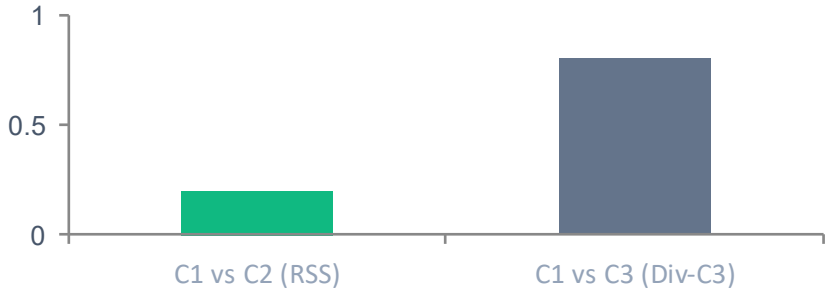
GPT-4o on OpenDialKG

given to 3 blind raters

Before the result, what should we expect?

How much did shuffling disrupt, relative to removing the relation entirely?

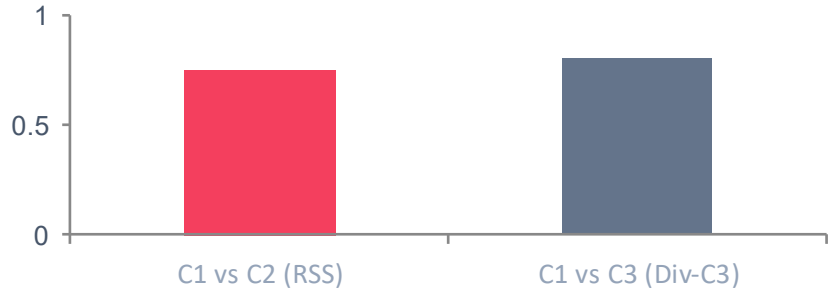
If models USE the relation



Shuffling (C2) causes much LESS disruption than removing (C3).

Ratio near 0 → GOD: the relation did real work.

If models IGNORE the relation



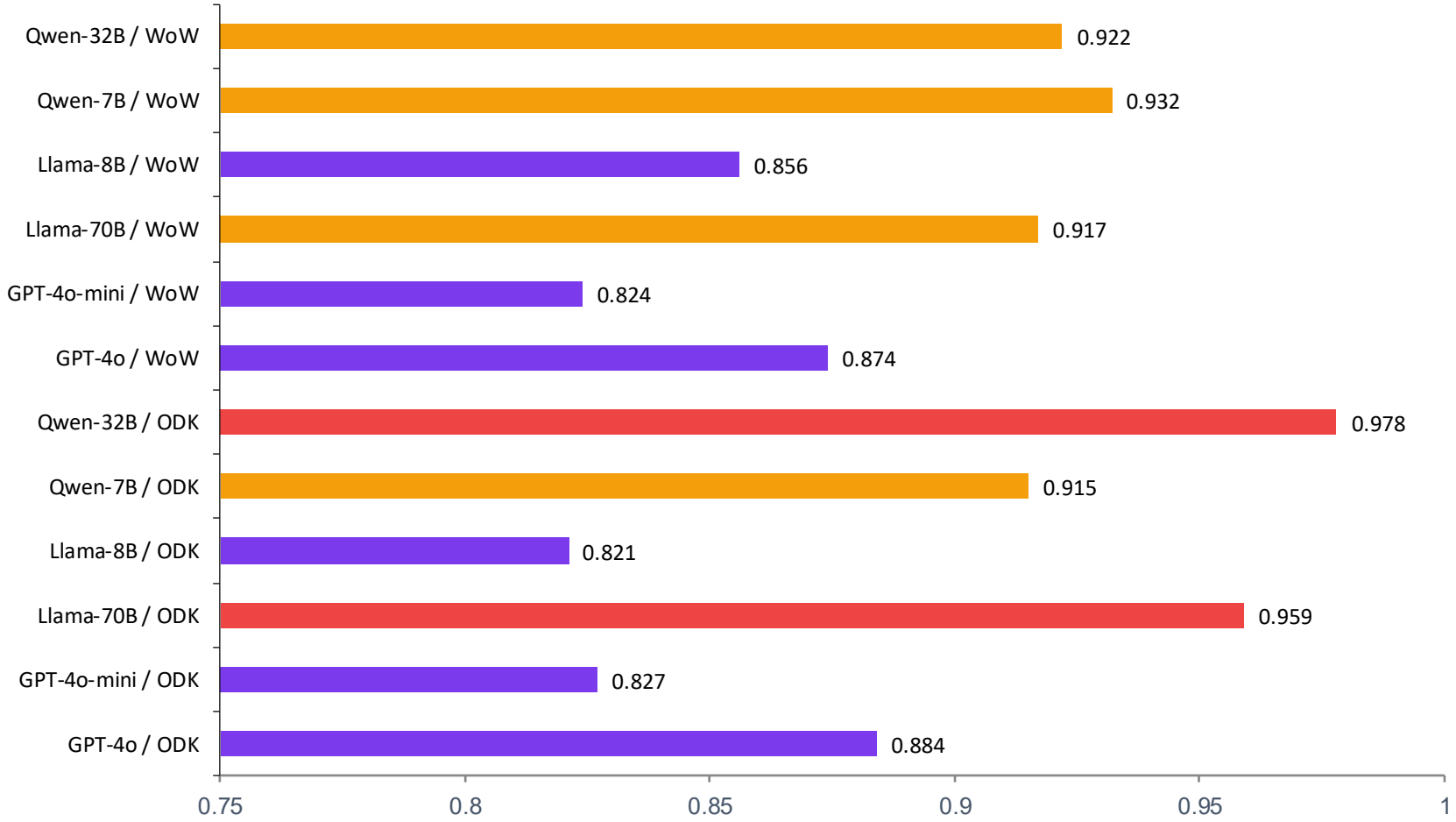
Shuffling (C2) causes SIMILAR disruption to removing (C3).

Ratio near 1 → BAD: the relation added almost nothing.

⚠ **Key direction:** Higher Ratio = worse. Ratio near 1.0 means the correct relation added almost nothing beyond the entities alone.

MAIN RESULT

Shuffling a relation changes the response almost as little as removing it entirely.



Ratio range

0.82–0.98

across all 12
model-dataset pairs

Wilcoxon $p < 0.001$
all 12 pairs

≥ 0.95 — extreme
0.90–0.95
< 0.90

STATISTICAL EVIDENCE

This is not noise.
It holds everywhere.

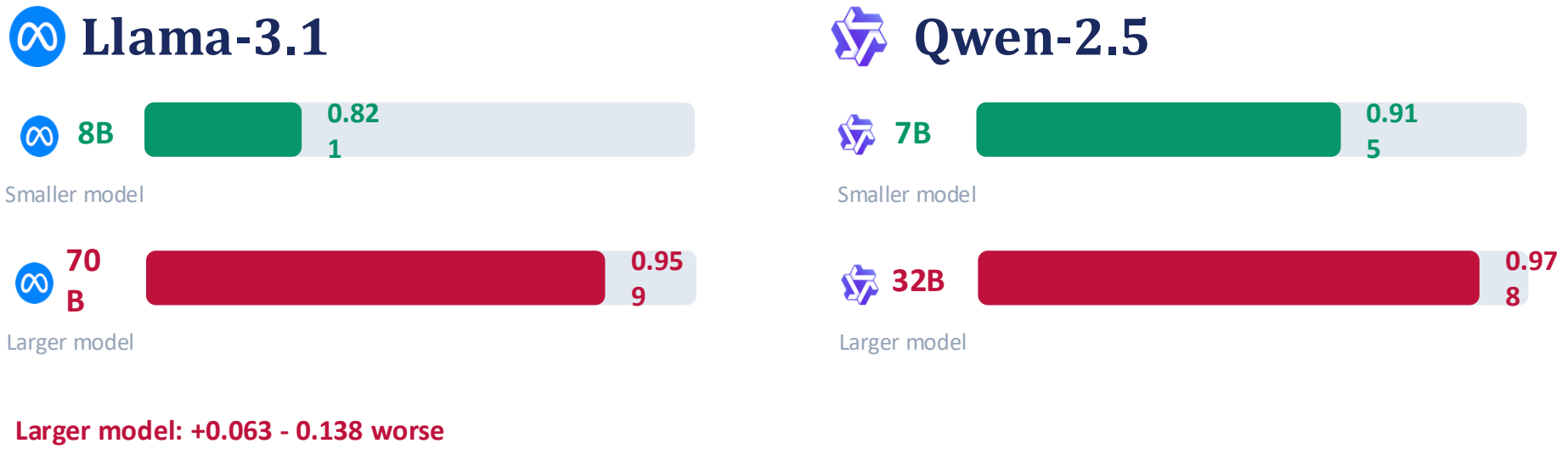
RSS(t) < Div_C3(t) for every single model, on every single dataset.

GPT-4o OpenDialKG p <0.001	**GPT-4o-mini** OpenDialKG p <0.001	**Llama-3.1-70B** OpenDialKG p <0.001	**Llama-3.1-8B** OpenDialKG p <0.001
GPT-4o WoW-KG p <0.001	**GPT-4o-mini** WoW-KG p <0.001	**Llama-3.1-70B** WoW-KG p <0.001	**Llama-3.1-8B** WoW-KG p <0.001
Qwen2.5-7B OpenDialKG p <0.001	**Qwen2.5-32B** OpenDialKG p <0.001	**Qwen2.5-7B** WoW-KG p <0.001	**Qwen2.5-32B** WoW-KG p <0.001

One-sided Wilcoxon signed-rank test: H_1: RSS(t) < Div_C3(t)

FINDING — SCALE

You might think bigger models fix it. They don't. They make it worse.



Why? Stronger parametric recall means the model leans even harder on what it already knows about the entity.

More world knowledge → more entity salience → relation label becomes even more irrelevant.

The Ratio distribution has two explanations.

1 Entity Salience

The model already knows enough about the entity from pretraining.

The relation becomes irrelevant, the response is entity-anchored regardless of which relation it receives.

Spearman ρ = −0.108 (OpenDialKG)

ρ = −0.071 (WoW-KG)

p < 0.001 · replicated across all 6 models

→ This explains low-RSS turns.

Model ignores the relation whether it's correct or wrong.

2 Surface-Form Confusion

The shuffled relation's wording carries real semantic content.

The model reads the label as natural Language, not as a typed ontological pointer, and follows the wording somewhere wrong.

[sport played] → Robert Duvall

→ Model pivots to his real baseball interest

RSS = 0.771 (factually true, contextually wrong)

→ This explains high-RSS turns.

But it's the wrong reason, wrong label hijacked the response.

EXAMPLES — WHAT THIS LOOKS LIKE

Let me show you both in action.

RSS ≈ 0 · Entity Salience

C1 Blood Promise – [has_genre] → young-adult fiction

"Blood Promise is a young adult fantasy novel, part of the Vampire Academy series by Richelle Mead."

C2 Blood Promise – [starred_actors] → young-adult fiction

"Blood Promise is a young adult fantasy novel, part of the Vampire Academy series by Richelle Mead."

Word-for-word identical. Entity salience took over.

The entity "Blood Promise" is domain-specific enough that relation label could not redirect the response.

RSS = 0.771 · Surface-Form Confusion

C1 The Apostle – [written_by] → Robert Duvall

"Robert Duvall both wrote and directed The Apostle, a 1997 drama in which he starred as a Pentecostal preacher."

C2 The Apostle – [sport played] → Robert Duvall

"Robert Duvall is known for his versatility as an actor. He has played many roles, including baseball-related ones…"

[sport played] sent the model to Duvall's baseball interest.

Factually true. Completely wrong for the context.
High RSS, but for the wrong reason.

HUMAN EVALUATION

We asked humans to verify.

100 turns · GPT-4o · OpenDialKG · 3 raters · blind to condition

53%

preferred the valid-KG response

vs only 8% shuffled

58%

couldn't tell which response was corrupted

95%

couldn't detect corruption on zero-RSS turns

Why we ran it:

RSS is automatic text comparison. A skeptic could say the responses just sound different but mean the same thing.

When entity salience dominates, the two responses are so similar that 95% of the time a careful human reader sees no difference. The wrong relation left no trace, because it was never used.

What we showed.

- **Relations are systematically underutilized.** Ratio 0.82–0.98. $p < 0.001$. 12 of 12 pairs.
- **Two mechanisms:** entity salience (low RSS) and surface-form confusion (high RSS). Neither is genuine relational reasoning.
- **Scale amplifies the problem.** Larger models lean harder on parametric recall; relation labels matter even less.
- **Humans confirm it.** 58% undetectable overall. 95% on zero-RSS turns. The wrong relation left no trace.
- **New tools:** WoW-KG (219 relation types, released) and RSS (reusable behavioral metric).

Current LLMs in KG-grounded dialogue mostly use entities, not typed relations.

Thank you.

Dominic Okonkwo · dominic.okonkwo@uga.edu

NeSLab · School of Computing · University of Georgia

Questions?