Do LLMs Use Relational Structure in KG-Grounded Dialogue? Evidence from Entity and Relation Perturbation

Dominic Okonkwo and Ismailcem Budak Arpinar
School of Computing
University of Georgia
{dominic.okonkwo, budak}@uga.edu

Abstract

Knowledge graph-grounded dialogue systems are built on the assumption that large language models leverage not only entity identity but also the typed relational structure connecting entities. We present the first controlled perturbation study to directly test this assumption, comparing model responses under three conditions: intact KG triples, relation-shuffled triples with entities preserved, and entity-only context. We evaluate six models across three families on two datasets, OpenDialKG and WoW-KG, a new corpus we construct by augmenting Wizard of Wikipedia with Wikidata triples. We introduce the Relation Sensitivity Score (RSS), a behavioral metric quantifying response divergence under relation perturbation. We find that RSS is significantly lower than the divergence from removing relations entirely across all twelve model-dataset combinations (Wilcoxon $p < 0.001$), with RSS-to-divergence ratios ranging from 0.821 to 0.978. Two mechanisms explain this pattern: entity salience, where the model already knows enough about the entity that the relation label becomes irrelevant (Spearman $\rho = -0.108$, $p < 0.001$), and surface-form confusion, where a shuffled label's wording alone pulls the response off course despite being invalid. Neither mechanism improves with model scale. We then asked human raters to confirm these findings, and the results agree: valid-KG responses are preferred in 53% of turns vs. 8% for shuffled-KG responses, yet raters cannot identify the corrupted condition in 58% of turns overall and 95% of low-sensitivity turns. These results show that relational structure is systematically underutilized in current KG-grounded dialogue under zero-shot prompting, with direct implications for how such systems are designed and evaluated.

1 Introduction

Knowledge graph (KG)-grounded dialogue systems are typically designed on one core assumption: that a model can use not just the entities in a knowledge graph, but the typed relations connecting them. This assumption shapes a wide range of design choices in KG-augmented dialogue, including path-based graph traversal in OpenDialKG (Moon et al., 2019), subgraph retrieval for response generation (Kang et al., 2023; Park et al., 2024), and relation-transition-aware modeling across dialogue turns (Wang et al., 2022; Tuan et al., 2022). Earlier methods take a simpler approach, adding external knowledge into dialogue models through neural conversation models and KG embeddings (Ghazvininejad et al., 2018; He et al., 2017). All of these methods build relation types directly into the model. But none of them test whether the model actually uses those relation labels when it generates a response.

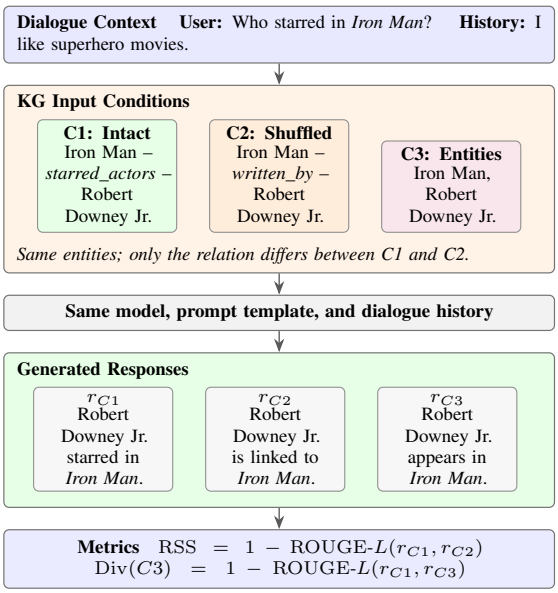

Figure 1: **Controlled perturbation setup.** We compare outputs under intact triples (C1), relation-shuffled triples (C2), and entity-only context (C3) to compute relation sensitivity (RSS) and divergence to the entity-only condition, Div(C3).

This leaves a basic question unanswered: when an LLM is given a KG triple during dialogue gener-

Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 375–384
August 2–5, 2026.

ation, does it really use the typed relation connecting the two entities, or does it just rely on the entities themselves? The answer matters. If relational structure is only weakly used, then the benefits we attribute to KG grounding may really just come from exposure to the right entities, not from any structured reasoning over relations (Longpre et al., 2021). If relation types are genuinely used, then modeling relational structure remains essential to system design. Either way, we need a direct test, not just an assumption.

In this work, we run that test. We conduct a controlled perturbation study to examine whether relational structure actually matters in KG-grounded dialogue generation. We build three input conditions: (C1) intact KG triples, (C2) relation-shuffled triples, where the entity pair stays the same but the relation label is randomly swapped, and (C3) entity-only context, with all relations removed. By holding the entities fixed and only changing the relation, we can isolate exactly what the relation label contributes to generation. We evaluate six models across three families, GPT-4o and GPT-4o-mini (OpenAI), Llama-3.1-8B and Llama-3.1-70B (Meta), and Qwen2.5-7B and Qwen2.5-32B (Alibaba), on two datasets: OpenDialKG (Moon et al., 2019) and WoW-KG, a new resource we build by augmenting Wizard of Wikipedia (Dinan et al., 2019) with Wikidata triples. To measure sensitivity to relational structure directly, we introduce the Relation Sensitivity Score (RSS): a metric that captures how much a model's output changes when the relation is perturbed.

Across every model and dataset we test, the same pattern shows up: responses under shuffled relations look much like responses under the correct relation. The change caused by shuffling a relation is consistently smaller than the change caused by removing relations altogether. We trace this pattern to two mechanisms: entity salience, where the model's response is driven by what it already knows about the entities, not the relation (Coca et al., 2023); and surface-form sensitivity, where the model reacts to the wording of the relation label rather than its actual structural role. Larger models do not fix this, which lines up with prior findings that factual grounding does not scale cleanly with model size (Mallen et al., 2023). In short, under zero-shot prompting, relational structure is consistently underused relative to entity identity. We are not claiming KG relations carry no information at all, only that their contribution at inference time is smaller than commonly assumed, which is exactly why controlled perturbation studies like this one matter.

Our contributions are as follows: (1) the first controlled perturbation study isolating the role of typed relation labels in LLM-based KG-grounded dialogue generation, evaluated across six models and two datasets; (2) WoW-KG, a new KG-grounded dialogue corpus constructed by augmenting Wizard of Wikipedia with Wikidata triples across 219 relation types; (3) the Relation Sensitivity Score (RSS), a reusable behavioral metric for evaluating structural sensitivity in KG-grounded generation; (4) evidence that entity salience and surface-form confusion jointly explain relational insensitivity, with the pattern replicated across all six models independently; and (5) the finding that model scale does not improve relational grounding, with larger models showing higher RSS-to-divergence ratios than smaller counterparts within the same family.

2 Related Work

KG-grounded dialogue generation. Prior work on KG-grounded dialogue assumes that relational structure contributes meaningfully to response generation. Moon et al. (2019) introduced OpenDialKG, a dialogue dataset grounded in explicit KG paths, along with a model that generates responses by reasoning over these paths. Later work builds on the same assumption in two main ways. One line retrieves and encodes structured subgraphs, through subgraph retrieval-augmented generation (Kang et al., 2023) and generative subgraph retrieval (Park et al., 2024). A second line models relational structure more directly, through large-scale KG traversal with interpretable reasoning paths (Tuan et al., 2022) and relation-transition-aware modeling across dialogue turns (Wang et al., 2022). Earlier methods take a simpler approach, adding external knowledge into dialogue models through neural conversation models and KG embeddings (Ghazvininejad et al., 2018; He et al., 2017). All of these works build relation types into the model design, but none of them directly test whether perturbing relational structure changes anything. Our work isolates this factor through controlled input manipulation.

LLMs and structured knowledge in dialogue. Recent studies examine how LLMs interact with structured knowledge in dialogue settings. Schneider et al. (2024) evaluate models on conversational

grounding tasks using knowledge graphs as a semantic layer between dialogue utterances and structured information, focusing on grounding act classification rather than knowledge use during response generation. We complement this line of work by studying whether relational structure affects generated responses under controlled perturbations.

Sensitivity to structured inputs. A related line of work investigates model sensitivity to variations in structured inputs. Coca et al. (2023) show that schema-guided dialogue state tracking is sensitive to schema wording and improve robustness by grounding the model in knowledge-seeking turns from the dialogue corpus in addition to the schema. In discourse parsing, Li et al. (2024) and Costa and Kosseim (2025) show that models can generate or predict discourse-relation structure when trained under supervised settings. Unlike these supervised settings, we focus on zero-shot generation and evaluate whether relational structure is used at inference time without additional training.

Parametric knowledge and entity salience. Prior work suggests that models often rely on memorized parametric knowledge even when contextual information is provided (Longpre et al., 2021), and that factual performance varies with entity popularity or frequency (Mallen et al., 2023). We extend this to KG-grounded dialogue by examining how relational sensitivity varies with entity salience.

Evaluation and datasets. Standard metrics for knowledge-grounded dialogue, such as BLEU, ROUGE-L, and Entity-F1, primarily capture lexical overlap and omit relational correctness (Papineni et al., 2002; Lin, 2004). None of these metrics can tell us whether a model actually used the relation in a KG triple, only whether its response overlaps with a reference. To close this gap, we introduce the Relation Sensitivity Score (RSS), which measures how much a response changes under relation perturbation. We also construct WoW-KG by augmenting Wizard of Wikipedia (Dinan et al., 2019) with Wikidata triples, giving us a second dataset with relation types of varying abstraction to evaluate against.

3 Methodology

3.1 Datasets

We need a dataset where every dialogue turn already has a real KG triple attached to it, so that we can build our three conditions, intact, shuffled, and entity-only, in a fair and consistent way. We use two datasets that satisfy this requirement differently: OpenDialKG, an existing KG-grounded dialogue benchmark, and WoW-KG, a new dataset we construct ourselves. This lets us test whether our findings hold across different knowledge sources and relation styles. Figure 1 illustrates our controlled perturbation setup.

OpenDialKG. OpenDialKG (Moon et al., 2019) is a benchmark corpus of human–human, multi-turn dialogues, and the standard dataset in prior KG-grounded dialogue work. Each dialogue is paired with explicit Freebase KG paths, so every assistant response already has a known, correct relation attached. The data were collected using a Wizard-of-Oz setup, in which a knowledge-aware assistant was shown candidate one- and two-hop facts from the KG at each turn, and wrote a response using the relevant ones. As a result, every assistant turn is linked to a grounding path of the form (entity, relation, entity), exactly the structure our framework needs.

The corpus spans four domains: movies, books, sports, and music. It is built on a filtered Freebase subgraph (Bollacker et al., 2008) with about 100,000 entities, 1.1 million facts, and 175 unique relation types. We use the standard 70/15/15 split, evaluating on 3,696 test turns, of which 3,216 remain after collision filtering.

WoW-KG. OpenDialKG alone is not enough. Its relation labels, such as *starred_actors* or *written_by*, are short and close to natural language, making them relatively easy to guess from wording alone. To test whether our findings generalize, we need a second dataset with more abstract relations. Wizard of Wikipedia (Dinan et al., 2019) is a large, well-known knowledge-grounded dialogue corpus, but it grounds dialogue in free-text Wikipedia passages, not structured triples, so it cannot be used directly in our framework. We build WoW-KG by adding real Wikidata triples on top of it, giving us the structured grounding we need while keeping its open-domain, encyclopedic style.

We construct WoW-KG in three steps. First, for each wizard turn, we apply spaCy named entity recognition (Honnibal et al., 2020) to the local dialogue context to identify the entities being discussed. Second, we query Wikidata via SPARQL (Prud'hommeaux and Seaborne, 2008) to retrieve all connected triples, in the form (entity, prop-

erty, value), keeping only properties with human-readable labels. Third, when multiple candidate triples are available, we keep the one whose subject is best supported by the associated Wikipedia passage, measured by surface-form overlap. This process yields 8,715 wizard turns, with 95.4% KG coverage and 219 unique relation types. WoW-KG relations are noticeably more abstract than OpenDialKG's: instead of short labels like *starred_actors* or *has_genre*, WoW-KG uses broader relations like *instance of*, *occupation*, and *country*. These are harder to guess from wording alone, making WoW-KG a stronger test of whether a model truly uses relational structure rather than pattern-matching familiar labels. We use the original Wizard of Wikipedia splits: 6,217 training, 1,531 validation, and 779 test turns, of which 778 remain after collision filtering.

The two datasets have very different collision rates: 13.0% for OpenDialKG versus 0.1% for WoW-KG. This follows from relation inventory size: OpenDialKG's smaller, more repetitive set of relations makes accidental matches more likely under random shuffling, while WoW-KG's much larger inventory makes this far less likely.

Attribute	OpenDialKG	WoW-KG
KG source	Freebase	Wikidata
# Relations	175	219
Train	14,908	6,217
Test	3,696	779
Test (filtered)	3,216	778
KG coverage	100%	95.4%
Collision rate	13.0%	0.1%

Table 1: **Dataset statistics.** Collision rate is the proportion of test turns excluded because relation shuffling produced the original relation by chance. KG coverage for WoW-KG is the proportion of wizard turns for which at least one Wikidata triple was successfully retrieved.

3.2 Input Conditions

We build three input conditions for every eligible turn in both datasets. Across all three, entity identity stays fixed, and only the relational layer changes. Table 2 shows all three conditions side by side, using the same entity pair throughout.

- **C1 (Full KG):** The original triple, with correct entities and correct relation label. This is standard KG-grounded dialogue generation.
- **C2 (Shuffled Relations):** The same entity pair, but with the relation label replaced by one drawn uniformly at random from the dataset's full relation inventory. This is our key perturbation condition: if the model's response barely changes, it tells us the typed relation label is not driving generation. We exclude turns where the shuffled relation happens to match the original (collision-filtered turns).
- **C3 (Entities Only):** Entity names only, with all relational structure removed. This gives us an entity-only baseline, so we can compare the effect of corrupting a relation against removing it entirely.

Condition	Input to model
C1 (Full KG)	*Iron Man –[starred_actors]–> Robert Downey Jr.*
C2 (Shuffled)	*Iron Man –[written_by]–> Robert Downey Jr.*
C3 (Entities Only)	*Iron Man, Robert Downey Jr.*

Table 2: **Example of the three input conditions applied to the same entity pair.** Entity identity is held fixed across all three conditions; only the relational layer varies.

When a turn has more than one available KG triple, we use a single annotated or selected triple. This keeps the perturbation targeted and the meaning of each condition unambiguous.

3.3 Models and Prompting

We evaluate six instruction-tuned models across three families. From OpenAI, we use GPT-4o and GPT-4o-mini, both accessed via API (OpenAI et al., 2024). From Meta, we use Llama-3.1-70B and Llama-3.1-8B (MetaAI et al., 2024); we access Llama-3.1-70B through Together.ai, and run Llama-3.1-8B locally. From Alibaba, we use Qwen2.5-7B and Qwen2.5-32B (Qwen et al., 2024). Llama-3.1-8B and both Qwen models run locally on NVIDIA A100 GPUs, using Hugging Face Transformers (Wolf et al., 2020).

We evaluate all models zero-shot, using one uniform prompt structure: a system instruction that tells the model to ground its response in the provided KG context, followed by the dialogue history and the condition-specific triple. We use the same template across every model and every condition, so that prompt variation never becomes a confound. We use no few-shot demonstrations and no fine-tuning. This setup tests one thing: whether pre-trained, instruction-following LLMs use relational structure when it is given to them at inference time, not whether they could be trained to use it.

3.4 Evaluation Metrics

Our primary metric is the Relation Sensitivity Score (RSS), which measures how much a model's response changes when only the relation label is perturbed. We compute RSS using ROUGE-L (Lin, 2004) between responses under different conditions. For turn t:

$$\mathrm{RSS}(t) = 1 - \text{ROUGE-L}\big(r_{C1}(t),\, r_{C2}(t)\big) \tag{1}$$

where $r_{C1}(t)$ and $r_{C2}(t)$ are the model responses under C1 and C2 respectively. RSS approaches 0 when the two responses are nearly identical, indicating insensitivity to the relation label, and increases as responses diverge. RSS is computed only on collision-filtered turns.

To contextualize RSS, we compute the divergence between the full-KG and entity-only responses:

$$\mathrm{Div}_{C3}(t) = 1 - \text{ROUGE-L}\big(r_{C1}(t),\, r_{C3}(t)\big) \tag{2}$$

This measures how much the response changes when relational structure is removed entirely. We then define the normalized ratio:

$$\mathrm{Ratio} = \frac{E\big[\mathrm{RSS}(t)\big]}{E\big[\mathrm{Div}_{C3}(t)\big]} \tag{3}$$

Here, the expectation is taken over collision-filtered test turns, for a given model and dataset. A Ratio near 1.0 tells us something important: shuffling the relation is nearly as disruptive as removing it altogether, which means the valid relation label adds little beyond entity identity alone.

We also report two standard metrics. ROUGE-L between C1 responses and gold reference turns gives us a general measure of generation quality. Entity F1 (EF1) gives us a proxy for factual grounding, based on token-level entity overlap. Neither metric tells us whether the relation label was used correctly, only whether the response looks lexically similar to a reference. This is exactly the gap RSS is designed to fill. For completeness, we also report BLEU-1 and BLEU-4 (Papineni et al., 2002) as additional lexical overlap metrics.

3.5 Statistical Testing

For each model–dataset pair, we test whether relation perturbation changes responses less than full relation removal using a one-sided Wilcoxon signed-rank test (Wilcoxon, 1945) on per-turn distributions:

$$H_1 : \mathrm{RSS}(t) < \mathrm{Div}_{C3}(t) \tag{4}$$

We also test whether per-turn RSS distributions differ between OpenDialKG and WoW-KG using a two-sided Mann–Whitney U test (Mann and Whitney, 1947) :

$$H_2 : \mathrm{RSS}_{\text{OpenDialKG}} \neq \mathrm{RSS}_{\text{WoW-KG}} \tag{5}$$

We use a two-sided test for H_2 because we have no prior reason to expect higher relational sensitivity on one dataset over the other. Together, these two tests let us separate three effects that are often tangled together in KG-grounded dialogue evaluation: the contribution of entity identity, the contribution of typed relation labels, and overall response quality.

4 Results and Discussion

4.1 Sensitivity to Relational Perturbation

Table 3 reports RSS, Div_{C3}, Ratio, ROUGE-L, and EF1 for all six models across both datasets using collision-filtered turns.

Across all twelve model–dataset combinations, RSS is consistently lower than Div_{C3}. The Ratio ranges from 0.821 to 0.978, indicating that corrupting the relation label produces a response change of similar magnitude to removing relational structure entirely. This pattern holds without exception and is consistent with the interpretation that entity identity, rather than the typed relation label, is the primary driver of response generation under this zero-shot prompting setup. We note that this does not imply relations are entirely unused, but their contribution appears limited relative to entities in this setting.

On OpenDialKG, the highest Ratios belong to Qwen2.5-32B (0.978) and Llama-3.1-70B (0.959), meaning these models treat a shuffled relation as nearly equivalent to the correct one. On WoW-KG, the same directional pattern holds across all models. GPT-4o-mini is a notable case: its RSS is higher on WoW-KG (0.481) than on OpenDialKG (0.395), which we attribute to Wikidata relation strings carrying stronger surface-level semantic content than Freebase labels, a point we discuss in Section 4.3. Entity F1 stays stable across all three conditions, which confirms that entity grounding is actively maintained even when the relation label is corrupted. We emphasize that these findings are specific to zero-shot prompting. Systems trained explicitly on relational reasoning tasks may behave differently, and we make no claim about supervised settings. Table 4 shows the per-turn RSS

Model	Dataset	n	RSS	Div_{C3}	Ratio	ROUGE-L	EF1	BLEU-1	BLEU-4
GPT-4o	OpenDialKG	3,216	0.544	0.616	0.884	0.221	0.152	0.156	0.031
GPT-4o-mini	OpenDialKG	3,214	0.395	0.477	0.827	0.202	0.131	0.142	0.023
Llama-3.1-70B	OpenDialKG	3,216	0.613	0.639	0.959	0.173	0.091	0.119	0.022
Llama-3.1-8B	OpenDialKG	3,216	0.563	0.686	0.821	0.186	0.117	0.130	0.023
Qwen2.5-7B	OpenDialKG	3,216	0.556	0.608	0.915	0.236	0.185	0.169	0.038
Qwen2.5-32B	OpenDialKG	3,216	0.580	0.593	0.978	0.225	0.164	0.162	0.031
GPT-4o	WoW-KG	778	0.522	0.597	0.874	0.123	0.220	0.110	0.011
GPT-4o-mini	WoW-KG	778	0.481	0.584	0.824	0.123	0.146	0.111	0.012
Llama-3.1-70B	WoW-KG	778	0.573	0.625	0.917	0.110	0.193	0.088	0.009
Llama-3.1-8B	WoW-KG	778	0.506	0.590	0.856	0.128	0.136	0.099	0.011
Qwen2.5-7B	WoW-KG	778	0.577	0.619	0.932	0.126	0.229	0.112	0.012
Qwen2.5-32B	WoW-KG	778	0.605	0.657	0.922	0.113	0.213	0.097	0.009

Table 3: **Main results.** RSS = mean Relation Sensitivity Score (higher = more sensitive to relation shuffling). Div_{C3} = mean response divergence when relational structure is removed entirely. Ratio = RSS / Div_{C3}. values near 1.0 indicate the valid relation label provides little additional guidance beyond entity identity. All turns are collision-filtered.

Model	Data	Low	Mid	High
GPT-4o	OpenDialKG	75	2,334	807
GPT-4o-mini	OpenDialKG	310	2,593	311
Llama-3.1-70B	OpenDialKG	24	2,114	1,078
Llama-3.1-8B	OpenDialKG	98	2,120	998
Qwen2.5-7B	OpenDialKG	89	2,245	882
Qwen2.5-32B	OpenDialKG	46	2,229	941
GPT-4o	WoW-KG	10	677	91
GPT-4o-mini	WoW-KG	37	583	158
Llama-3.1-70B	WoW-KG	0	661	117
Llama-3.1-8B	WoW-KG	60	507	211
Qwen2.5-7B	WoW-KG	6	599	173
Qwen2.5-32B	WoW-KG	1	561	216

Table 4: **Per-turn RSS distribution.** Counts of turns with Low RSS (< 0.05; virtually insensitive to relation shuffling), Mid RSS ($0.05 \leq \text{RSS} < 0.70$; moderate sensitivity), and High RSS (≥ 0.70; shuffled labels substantially redirected the response). Low, Mid, and High are mutually exclusive and sum to n in every row, matching the totals reported in Table 3.

distribution across all models and datasets. This distribution reveals substantial variation, which we examine further in Sections 4.2 and 4.3.

4.2 Statistical Evidence

Table 5 reports statistical test results. The Wilcoxon tests are significant at $p < 0.001$ for all twelve model–dataset combinations, confirming that RSS is significantly lower than Div_{C3} universally, not just on average. The dominance of entity identity over the typed relation label is not model-specific or dataset-specific, it holds across every configuration evaluated. The Mann–Whitney tests tell a more nuanced story: four of six models show significant RSS differences between OpenDialKG and WoW-KG, but the direction is not uniform. This indicates that dataset characteristics such as relation inventory size and abstraction level modulate relational sensitivity for some models without producing a consistent directional effect.

Model	Wilcoxon p	Mann–Whitney p
GPT-4o	<0.001 (***)	<0.001 (***)
GPT-4o-mini	<0.001 (***)	<0.001 (***)
Llama-3.1-70B	<0.001 (***)	<0.001 (***)
Llama-3.1-8B	<0.001 (***)	0.454 (ns)
Qwen2.5-7B	<0.001 (***)	0.374 (ns)
Qwen2.5-32B	<0.001 (***)	0.014 (*)

Table 5: **Statistical tests.** Wilcoxon signed-rank test (one-sided) evaluates H_1: $\text{RSS}(t) < \text{Div}_{C3}(t)$. Mann–Whitney U test (two-sided) evaluates H_2: $\text{RSS}_{\text{OpenDialKG}} \neq \text{RSS}_{\text{WoW-KG}}$. Significance: *** $p < 0.001$, * $p < 0.05$, ns $p \geq 0.05$.

The most striking entry in Table 4 is Llama-3.1-70B on WoW-KG, which produces zero turns with RSS < 0.05, it never completely ignores relation shuffling on this dataset. We attribute this to Wikidata relation strings being lexically richer than Freebase labels, making the model more consistently reactive to their surface content.

4.3 Entity Salience and Surface-Form Confusion

Table 4 shows that the per-turn RSS distribution is far from uniform. A meaningful subset of turns has RSS < 0.05, meaning the model is essentially in-

[0]For GPT-4o-mini on OpenDialKG, 2 of 3,216 collision-filtered turns produced an empty response and were excluded from RSS computation, yielding $n = 3{,}214$.

Type	KG context ($C_1 \rightarrow C_2$)	C_1 response	C_2 response
Zero-RSS	Blood Promise *C1:* [has_genre] → young-adult fiction *C2:* [starred_actors] → young-adult fiction	Blood Promise is a young adult fantasy novel and part of the Vampire Academy series by Richelle Mead.	Blood Promise is a young adult fantasy novel and part of the Vampire Academy series by Richelle Mead.
Zero-RSS	Dale Carnegie *C1:* [written_by] → How to Win Friends *C2:* [directed_by] → How to Win Friends	Dale Carnegie wrote *How to Win Friends and Influence People*, one of the best-selling self-help books ever published.	Dale Carnegie wrote *How to Win Friends and Influence People*, one of the best-selling self-help books ever published.
High-RSS (0.771))	The Apostle *C1:* [written_by] → Robert Duvall *C2:* [sport played] → Robert Duvall	Robert Duvall both wrote and directed *The Apostle*, a 1997 drama in which he starred as a Pentecostal preacher.	Robert Duvall is known for his versatility as an actor. He has played many roles, including baseball-related ones, and is known for his passion for the sport.
High-RSS (0.814))	Bruce Almighty *C1:* [starred_actors] → Jim Carrey *C2:* [directed_by] → Jim Carrey	Jim Carrey starred in *Bruce Almighty*, playing a frustrated TV reporter who is granted divine powers.	Jim Carrey is well known for his comedic roles. He has worked with many directors throughout his career.
High-RSS (0.714)	LotR *C1:* [has_genre] → Fantasy *C2:* [sport played] → Fantasy	*The Lord of the Rings* is an epic high-fantasy novel by J. R. R. Tolkien, defining the genre.	Fantasy is also a term used in sports, particularly in fantasy sports leagues where participants manage virtual teams.

Table 6: Qualitative examples from GPT-4o-mini on OpenDialKG. Zero-RSS turns show minimal change in responses despite relation perturbation. High-RSS turns show cases where the shuffled relation semantically redirects generation. All turns are collision-filtered. Entity pairs are identical across C_1 and C_2; only the relation label changes.

Metric	OpenDialKG	WoW-KG
Low RSS	15.7 ± 8.2	14.7
High RSS	13.4 ± 6.5	11.4
ρ	-0.108	-0.071

Table 7: **Entity salience analysis.** Mean entity character length by RSS category and Spearman ρ between entity name length and per-turn RSS. All differences are statistically significant (Mann–Whitney $p < 0.001$). The correlation is consistent across all six models (ρ range: -0.099 to -0.131).

sensitive to relation shuffling. Another substantial fraction exceeds 0.70. This is not noise; it reflects two distinct behavioral patterns, which we examine in turn.

The first pattern is entity salience. This happens on turns where the head entity activates strong parametric knowledge from pretraining. In these cases, the model generates an entity-anchored response no matter which relation label is attached. To test this, we compute the Spearman rank correlation (Spearman, 1904) between head entity name length, used here as a proxy for entity familiarity, and per-turn RSS, across all model–turn pairs. Results are in Table 7.

The correlation is negative and significant on both datasets ($p < 0.001$), and the pattern repeats across all six models independently. Zero-RSS turns have significantly longer entity names than high-RSS turns: 15.7 vs. 13.4 characters on OpenDialKG, and 14.7 vs. 11.4 on WoW-KG. Longer names tend to belong to more domain-specific, less universally familiar entities, for example *El amor en los tiempos del cólera* or *Vampire Academy: The Graphic Novel*. For these entities, the entity itself constrains the topic so tightly that no relation label can redirect the response. When entity salience is high enough, the relational path gets bypassed entirely. This is a specific failure mode of KG-grounded generation.

The second pattern is surface-form confusion. This is the complementary behavior we see in high-RSS turns. Here, the shuffled relation label carries enough surface-level semantic content to redirect the response, not because the model is reasoning correctly over KG structure, but because it reacts to the wording of the corrupted string. Table 6 gives representative examples of both patterns.

Two examples illustrate this clearly. In the Apostle example, the model pivots to Robert Duvall's real interest in baseball, a true biographical fact, but one that has nothing to do with the original

dialogue context. In the Lord of the Rings example, the shuffle produces *–[Sport played]–> Fantasy*, and the model interprets *Fantasy* as fantasy sports rather than the literary genre. In both cases, the model treats the relation string as a piece of natural language to react to, not as a typed pointer inside a formal ontology. This same mechanism explains the negative Spearman correlation above. Short, high-salience entities such as *Batman* and *Comedy* often pair with shuffled labels whose wording is vivid enough to redirect the response. The result is high RSS, but it comes from lexical reactivity, not genuine relational understanding.

Together, these two patterns explain the full distribution in Table 4. Zero-RSS turns are entity-driven: the relation label is redundant, because the model already knows what to say about the entity. High-RSS turns are relation-string-driven: a vivid, corrupted label pulls the response off course. Most turns fall somewhere between these two extremes. In these cases, both entity identity and the surface wording of the relation play a partial role, but neither reflects genuine use of the KG's typed relational structure.

4.4 Effect of Model Scale

Larger models do not show lower RSS. Within Llama-3.1, the 70B model (RSS = 0.613 on OpenDialKG) exceeds the 8B model (0.563). Within Qwen2.5, the 32B model (0.580) exceeds the 7B model (0.556) on OpenDialKG and similarly on WoW-KG. Qwen2.5-32B achieves the highest Ratio overall (0.978), meaning the largest Qwen model is the most likely of all evaluated models to treat a shuffled relation as equivalent to the correct one.

A plausible interpretation is that stronger parametric recall in larger models amplifies the entity salience effect rather than improving structural sensitivity, but we treat this cautiously given that RSS also reflects general response variability.

4.5 Human Validation

To check whether RSS reflects differences a human would actually notice, we ran a human evaluation on 100 stratified turns from GPT-4o on OpenDialKG (34 zero-RSS, 33 mid-RSS, 33 high-RSS). Three raters scored each response pair independently, blind to which condition produced which response. They rated factual correctness (Q1, 1–3), relation use and misleading (Q2, 0–2), fluency (Q3, 1–3), overall preference (Q4: A / SAME / B), and whether they could tell which response used the corrupted relation (Q5: A / B / CANNOT_TELL). Results are in Table 8.

Group	Q1-A	Q1-B	Q2-A	Q2-B
All	2.88	2.52	1.87	0.53
Low	2.95	2.90	1.95	0.05
Mid	2.85	2.48	1.82	0.73
High	2.82	2.18	1.85	0.85
Group	**Q3-A**	**Q3-B**	**Q4 (A/S/B)**	**Q5**
All	2.98	2.64	53/39/8	58%
Low	3.00	2.95	10/90/0	95%
Mid	2.97	2.52	73/18/9	42%
High	2.97	2.45	82/3/15	33%

Table 8: **Human evaluation results across 100 stratified turns.** Scores are reported overall and by RSS group (Low, Mid, High). A denotes the response generated with valid KG context; B denotes the response generated with a shuffled relation. Q1 and Q3 are rated on a 1–3 scale, where higher is better. Q2-A is rated on a 0–2 scale, where higher indicates clearer use of the valid relation. Q2-B is rated on a 0–2 scale, where higher indicates that the model was more misled by the shuffled relation. Q4 reports preference percentages (A/S/B). Q5 reports the percentage of turns where raters could not identify which response used the corrupted relation.

Valid-KG responses score higher on factual correctness (2.88 vs. 2.52) and fluency (2.98 vs. 2.64) than shuffled-KG responses. Both gaps grow wider across RSS strata. On high-RSS turns, the factual correctness gap reaches 0.64 points (2.82 vs. 2.18), which fits with surface-form confusion degrading response quality whenever the corrupted label pulls the response off course. Raters preferred the valid-KG response in 53% of turns, versus only 8% for the shuffled response, with the remaining 39% rated as equivalent.

The most telling result is Q5. Raters could not tell which response used the corrupted relation in 58% of turns overall, and this number rises to 95% on zero-RSS turns. When entity salience dominates, the two responses look the same even to a careful human reader. This is direct human evidence that the relation label was not driving the response in the first place. Response A is clearly rated as using the valid relation in 88% of turns. Response B, by contrast, is rated as clearly or partially misled in 40% of turns overall, and this rises further on high-RSS turns (Q2-B = 0.85), which supports the surface-form confusion account. On zero-RSS turns, the Q2-B score is near zero (0.05).

This is expected: when the model ignores the relation label entirely, the shuffled relation leaves no trace in the response, so raters correctly score B as not misled. Across every dimension we measure, the human evaluation lines up with the automatic RSS findings.

5 Conclusion

We presented the first controlled perturbation study of relational structure in KG-grounded dialogue generation. Across six models and two datasets, RSS-to-divergence ratios ranging from 0.821 to 0.978 demonstrate that typed relation labels are systematically underutilized relative to entity identity under zero-shot prompting, a finding confirmed by Wilcoxon tests significant at $p < 0.001$ across all twelve model–dataset combinations.

Two mechanisms explain this pattern. Entity salience describes turns where strong parametric knowledge of the head entity renders the relation label redundant (Spearman $\rho = -0.108$ on OpenDialKG, $\rho = -0.071$ on WoW-KG, both $p < 0.001$, replicated across all six models). Surface-form confusion describes turns where a semantically provocative shuffled relation string redirects the response through lexical reactivity rather than structural reasoning. Model scale does not improve relational grounding; larger models show higher ratios than smaller counterparts within both the Llama-3.1 and Qwen2.5 families.

Human evaluation confirms these findings: valid-KG responses are preferred in 53% of turns vs. 8% for shuffled responses, yet raters cannot identify the corrupted condition in 58% of turns overall and 95% of low-sensitivity turns. We do not argue that KG grounding is without value; stable EF1 scores confirm that entity exposure improves factual grounding. Rather, the specific contribution of relation labels has been over-attributed. We encourage future work to evaluate KG-grounded systems against relation-perturbed baselines as standard practice.

6 Limitations

This study evaluates zero-shot generation only; models trained explicitly on relational reasoning may behave differently. Head entity name length is a coarse proxy for entity salience; more precise alternatives include entity mention frequency in pretraining corpora, number of incoming Wikipedia links, or KG node degree. Human evaluation covers GPT-4o on OpenDialKG only. Domain-level RSS breakdown within OpenDialKG is left to future work, as domain labels are not recoverable from the available metadata. Finally, our findings are specific to open-domain dialogue with well-known entities. In low-resource or proprietary settings, such as medical NLG where relations hold between patients, medicines, and procedures, parametric knowledge is insufficient and relational structure may be indispensable. Extending this evaluation to such domains is an important direction for future work.

References

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, page 1247–1250, New York, NY, USA. Association for Computing Machinery.

Alexandru Coca, Bo-Hsiang Tseng, Jinghong Chen, Weizhe Lin, Weixuan Zhang, Tisha Anders, and Bill Byrne. 2023. Grounding description-driven dialogue state trackers with knowledge-seeking turns. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 444–456, Prague, Czechia. Association for Computational Linguistics.

Nelson Filipe Costa and Leila Kosseim. 2025. Multilingual implicit discourse relation recognition with multi-label hierarchical learning. pages 48–61.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5110–5117. AAAI Press.

He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. 2017. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1766–1776, Vancouver, Canada. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python. *Zenodo*.

Minki Kang, Jin Kwak, Jinheon Baek, and Sung Hwang. 2023. Knowledge graph-augmented language models for knowledge-grounded dialogue generation.

Chuyuan Li, Yuwei Yin, and Giuseppe Carenini. 2024. Dialogue discourse parsing as generation: A sequence-to-sequence LLM-based approach. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 1–14, Kyoto, Japan. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. pages 7052–7063.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. pages 9802–9822.

Henry B. Mann and Douglas R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18:50–60.

MetaAI, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, and 511 others. 2024. The llama 3 herd of models.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Opendialkg: Explainable conversational reasoning with attention over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 181 others. 2024. GPT-4o System Card. *arXiv e-prints*, arXiv:2410.21276.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Jinyoung Park, Minseok Joo, Joo-Kyung Kim, and Hyunwoo J. Kim. 2024. Generative subgraph retrieval for knowledge graph–grounded dialog generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21167–21182, Miami, Florida, USA. Association for Computational Linguistics.

Eric Prud'hommeaux and A Seaborne. 2008. Sparql query language for rdf, w3c recommendation.

Qwen, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, and 25 others. 2024. Qwen2.5 technical report. *ArXiv*, abs/2412.15115.

Phillip Schneider, Nektarios Machner, Kristiina Jokinen, and Florian Matthes. 2024. Bridging information gaps in dialogues with grounded exchanges using knowledge graphs. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 110–120, Kyoto, Japan. Association for Computational Linguistics.

Charles Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Yi-Lin Tuan, Sajjad Beygi, Maryam Fazel-Zarandi, Qiaozi Gao, Alessandra Cervone, and William Yang Wang. 2022. Towards large-scale interpretable knowledge graph reasoning for dialogue systems. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 383–395, Dublin, Ireland. Association for Computational Linguistics.

Kexin Wang, Zhixu Li, Jiaan Wang, Jianfeng Qu, Ying He, An Liu, and Lei Zhao. 2022. Rt-kgd: Relation transition aware knowledge-grounded dialogue generation. In *The Semantic Web – ISWC 2022: 21st International Semantic Web Conference, Virtual Event, October 23–27, 2022, Proceedings*, page 319–335, Berlin, Heidelberg. Springer-Verlag.

Frank. Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics*, 1:196–202.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Scao, Sylvain Gugger, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. pages 38–45.