

Detecting Large Language Model Hallucinations with Knowledge-Graph Integrity Constraints: An Evaluation Framework Using Wikidata

Simon D. Herbst^{1,*}, Miguel Vazquez^{1,*}

¹Vienna University of Economics and Business (WU Vienna), Welthandelsplatz 1, 1020 Vienna, Austria

Abstract

Knowledge graphs encode not only factual statements but also integrity constraints that specify in machine-readable form the conditions under which a statement is valid. As our key contribution, we show that such constraints can reject a factual statement without requiring a reference answer, a source document, or a language model as judge. We operationalize this capability in Wikidata as a post-generation audit of unaided LLM output. Each answer is grounded in an entity of the graph, completed into a statement, and tested against the constraints declared for its property. We evaluate 7,264 questions spanning four LLMs (3B–284B parameters), six properties, two constraint types, and two prompt framings, over 204 historical persons born between 1700 and 1900, and an additional set of 101 politicians for one property. Of the 6,142 evaluable answers, 891 (14.5%) violate a constraint, a figure that should be interpreted as a conservative indicator rather than a comprehensive measure of hallucination. Larger LLMs are less likely to assert a link to an entity unrelated to the subject, and stricter prompts are less likely to elicit an entity of an impermissible type. Neither larger LLMs nor stricter prompts reduce boundary errors: existing, closely related entities that lie outside the required temporal or type range.

Keywords

Large Language Models, Knowledge Graphs, Wikidata, Integrity Constraints, Hallucination Detection

1. Introduction

While Knowledge Graphs (KGs) structure factual knowledge into explicit, machine-readable relations [1], Large Language Models (LLMs) hold their knowledge implicitly, in the weights of a fundamentally probabilistic model. Because they generate responses by predicting likely token sequences without verifying the factual consistency of each claim, they can output plausible falsehoods, which are commonly termed hallucinations [2]. The problem persists even as LLMs are increasingly deployed in domain-critical applications requiring high factual fidelity [3, 4, 5].

Detecting a factual hallucination does not necessarily require knowledge of the correct answer. Some generated claims can be rejected because they violate integrity constraints, the necessary conditions a knowledge graph such as Wikidata encodes for the validity of its statements [6]. Consider, for example, the claim that Carl Maria von Weber (Q154812) was educated at the Music and Arts University of the City of Vienna (Q876640). Both the subject and the object correspond to valid Wikidata entities. However, the property *educated at* (P69) carries a contemporary constraint, which the claim violates: the university was founded after Weber’s death. The claim can therefore be rejected as temporally inconsistent without establishing where Weber was in fact educated.

This example motivates the use of KG constraints as a post-generation validation layer. Unlike reference-based evaluation, constraint validation can reject certain claims without requiring the correct answer, external source documents, or another language model as judge. As a focused initial study, we investigate this approach in Wikidata by examining statements about historical persons and two constraint types: value-type and contemporary constraints. We ask:

Wikidata’26: Wikidata Workshop at ISWC 2026

*Corresponding author.

✉ simon.david.herbst@s.wu.ac.at (S. D. Herbst); miguel.vazquez@wu.ac.at (M. Vazquez)

id 0009-0001-0456-1871 (S. D. Herbst); 0009-0002-9067-959X (M. Vazquez)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

To what extent, and in which ways, do factual statements generated by LLMs violate selected Wikidata integrity constraints?

To answer this question, we construct a controlled evaluation pipeline in which LLM-generated answers are mapped to Wikidata entities and converted into candidate statements. We query four LLMs under two prompt formulations for the values of six properties concerning 204 historical persons born between 1700 and 1900 (and an additional set of 101 politicians), and test the resulting statements against the two constraint types declared for those properties. Because constraint validation is only possible when a generated answer can be linked to a Wikidata entity, we report entity-linking coverage separately from constraint validity.

This paper makes three contributions. First, we operationalize Wikidata property constraints as a validation mechanism for LLM-generated factual statements. Second, we provide a reproducible pipeline and an accompanying dataset for assessing constraint satisfaction. Third, we present an empirical analysis of the prevalence of constraint violations and of how their frequency and character vary across LLMs, prompts, and properties. We release the implementation and resulting evaluation data to support further research ¹.

Across 6,142 evaluable answers, 891 14.5% violate at least one of the selected constraints. This indicates that Wikidata constraints can automatically detect a non-trivial subset of factual errors without reference answers. Since satisfying these constraints does not guarantee correctness, the violation rate should be interpreted as a conservative indicator rather than a comprehensive measure of hallucination.

2. Related Work

Combining LLMs and knowledge graphs allows researchers to exploit the complementarity between the structural knowledge of KGs and the implicit knowledge of LLMs [7]. Two main lines of this work address hallucination: approaches that augment generation with knowledge-graph content, and benchmarks that use a knowledge graph to evaluate what a model has already generated. Below, we examine how recent methods operationalize both paradigm directions.

The first line, in which the graph serves as an external knowledge source at inference time, has received the most attention. With GraphRAG, Edge et al. build a local KG from a text corpus [8]. It is clustered into hierarchical communities with pre-generated summaries, which then support LLMs in question answering. The KG serves as a structural index over that corpus, not as an external graph such as Wikidata. Pusch and Conrad prompt an LLM to re-write a medical question into a query for a KG, and then augment its generation with the query result [9]. Darari et al. apply the same idea to counting questions on Wikidata, in a prior edition of this workshop [10]. In one of their variants the model generates SPARQL over the counting properties Wikidata records. In the other, the triples of the entity in question are verbalized and placed in the model’s context. In the approaches discussed so far the graph enters through the model’s context, whether as retrieved summaries, as query results, or as verbalized triples. Luo et al. instead intervene during token generation: Graph-Constrained Reasoning influences the decoding phase directly, so that every generated reasoning path is grounded in the KG [11]. Surveying this line as a whole, Agrawal et al. report evidence that KG augmentation can substantially improve the factual reliability of LLMs [12]. Across this line the graph supplies content, and the object of measurement is how effectively a model exploits the context and tools provided to it, rather than the factual knowledge encoded in its parameters.

In the second line the graph does not participate in generation. The model answers unaided, and the graph is applied afterwards, either to construct the questions put to it or to score the answers it returns. Robertson et al. introduce KGHaluBench, a benchmark that constructs complex, open-ended questions using a Knowledge Graph [13]. It prompts the LLM for an overview of an entity and measures conceptual hallucinations and factual errors. Here, however, the graph only sources and scores the

¹<https://github.com/eddievanhalensolo/kg-halu-audit>

questions: KGHaluBench measures recognition and factual depth, not whether an answer respects the graph’s own integrity constraints. Sansford et al. evaluate hallucination in closed-domain tasks, comparing answers not against world knowledge but solely against the context document [14]. Their framework, GraphEval, prompts the model to decompose its answer into KG triples and checks each with a Natural Language Inference model, flagging those likely to contradict the document. Lee et al. generalize integrity constraints to the processing of LLM outputs [15]. A user declares constraints the answer must satisfy, such as regex matching or logical soundness, and a query optimizer decides whether to enforce them during decoding or post-generation. This employs integrity constraints but not KGs, and the semantic constraints are verified primarily by fact-checking models and secondary LLMs. In the studies that do examine adherence to a graph, its integrity constraints are seldom the evaluative instrument; adherence is assessed through NLI models or secondary LLM judges instead.

While prior work primarily employs knowledge graphs either to augment model context during generation or to provide ground-truth answers for benchmarking, these approaches do not assess the inherent structural consistency of unaided LLM outputs. Constraints are well established as an instrument of graph quality control, but they have rarely been directed at the factual and schema-level boundaries an LLM respects unaided. That is the gap this paper addresses: we have language models answer unaided and test the resulting statements against Wikidata’s integrity constraints, asking to what extent, and in which ways, those answers violate them.

3. Approach

The audit runs in three steps. We ask an LLM for a single fact, with the question specifying a subject and a property, so that only the value is left to the model. That value is then grounded in an entity of the graph, and the completed triple is tested against the constraints declared for the property. The procedure depends on three conditions: the constraints must be machine-readable, the subject must carry the metadata the constraint requires, and the answer must be sufficiently specific to resolve to a single entity. Wikidata meets the first; the setup below is built to meet the other two. We address these in turn: the two constraints, the subjects and properties we ask about, the LLMs and prompt framings employed, the path from prose to an entity, and the rules that determine a violation. Parameters relevant only to reproduction are documented in the released pipeline; we report here those that affect how the results should be interpreted.

A value-type constraint limits what a value entity may be an instance or a subclass of, which is also called the range of a property [16]. The range of country of citizenship (P27) comprises nation, city-state, empire, and similar classes. Each constraint also declares the relation through which its check is evaluated, either *instance of* alone, or *instance or subclass of*, which admits a transitive subclass. The contemporary constraint requires that the lifespans of the two linked entities overlap in time [17]; this is the constraint violated by the Carl Maria von Weber example above.

Our subjects are 204 humans born between 1700 and 1900, drawn from a Wikidata snapshot of 4 June 2026. A well-randomized set of one entity class keeps the property space tractable while remaining sufficiently generalizable for a preliminary study. Historical figures are almost always recorded with the birth and death dates the contemporary check requires; we retain only those individuals who have both complete lifespan dates and a recorded gender (P21).

The set is stratified by popularity, occupation, and gender. Popularity is measured by sitelinks: we keep two strata, obscure at 10 to 25 and popular at 75 or more, discarding the middle band, an extreme-groups design [18]. Occupation is divided into arts, science, and writing. Within each stratum, candidates are ranked by how many of the evaluated properties they record, so that a subject with several outranks one with a single property. Gender is sampled in alternation, one woman then one man, so the set does not inherit the pool’s male skew. The alternation largely holds in the obscure stratum (46 women to 56 men) but breaks down in the popular one (22 to 80). The full set comprises 68 women and 136 men. One filter precedes these: candidates whose recorded citizenship is already a contemporary violation are excluded, so the reported P27 rate is not an artifact of the sample.

We check the two constraint types across six properties, so that results hold at the level of the constraint type rather than of a single idiosyncratic property. All six carry a value-type constraint, four of them a contemporary one.

Property	Label	Value-type	Contemporary
P27	country of citizenship	Yes	Yes
P69	educated at	Yes	Yes
P39	position held	Yes	No
P166	award received	Yes	No
P463	member of	Yes	Yes
P102	member of political party	Yes	Yes

Table 1

The six properties and the constraint types checked on each. P102 is evaluated on its own politician-focused subject set.

The two gaps reflect what Wikidata defines and records rather than a decision to omit checks. Award received (P166) has no contemporary constraint defined in Wikidata, and position held (P39) yields answers whose entities frequently lack the inception date (P571) the contemporary check requires. Every property is recorded on at least 113 of the 204 subjects, except member of political party (P102), which only 31 of them carry. We therefore constructed a separate set of 101 politicians for P102, whose figures accordingly describe a different population.

The questions were posed to four LLMs, spanning 3B to 284B total parameters: Ministral 3B, Gemma 31B, Qwen 122B, and DeepSeek Flash at 284B. The two smaller LLMs are dense, while Qwen and DeepSeek Flash are mixture-of-experts models that activate approximately 10B and 13B parameters per token. All four were run with reasoning mode disabled.

The exact wording of a prompt influences the result [19], so every property is queried in two framings: the strict framing remains as close as possible to the triple of interest, while the colloquial framing expresses the same request in less precise terms, e.g. “What was Arthur Schopenhauer’s country of citizenship?” against “What was Arthur Schopenhauer’s citizenship?” Where a property admits several values, the prompt requests a single one; multi-valued answers are out of scope, and we therefore evaluate one value per question.

A model answer is free prose, so before any constraint can be evaluated we reduce it to the single string that answers the question and link that string to a Wikidata entity. Extraction proceeds in layers, applied from specific to general. A refusal guard is applied first and discards answers containing no checkable content, such as “I do not have a record of that”; hedged answers are retained, since “I am not sure, but he won the Nobel Prize” still asserts a claim. Next, an explicit “Answer:” or “Conclusion:” cue selects the bold span that follows it. Failing that, the layers match the relation cues of the property in question, e.g. “citizen of” for country of citizenship (P27), “won the” for award received (P166), “studied at” for educated at (P69), and select the emphasised span adjacent to the cue, whether preceding or following it: “won the **Nobel Prize**” gives Nobel Prize, “was a **French** citizen” gives French. The same cues are then applied once more without the emphasis requirement, in order to capture unbolded answers. As a final fallback, the first bold span is taken. The resulting string is constrained to between 3 and 120 characters, containing no commas and no bare numerals, and for P27, where LLMs often answer with an adjective, a lookup table maps “German” to Germany (Q183).

The mapping from string to Wikidata entity proceeds as a cascade. We normalize it, stripping parentheses, keeping the text before a comma and dropping leading articles, then search Wikidata. When a search returns no result we retry with spelling variants, then strip a trailing location, and finally fall back to CirrusSearch, which matches full article text rather than labels and aliases. Every result is then filtered against a hand-collected set of 45 noise classes, since a string may match an unrelated entity bearing the same name: “Austria” also matches the song “I am from Austria”. The surviving candidate yields a QID, whose instance of (P31) classes are retrieved for the check.

The two constraints are then checked by different mechanisms. For the value-type check we read the

retrieved P31 classes; an entity carrying none is recorded as lacking metadata and is excluded from the evaluable set. The answer passes if at least one of them is a permitted class, or, where the constraint declares *instance or subclass of*, a transitive subclass of one, which we test against precomputed closures. Member of political party (P102) is the exception: a named party is typically an instance of a party subclass such as *communist party*, not of *political party* itself, so we check it against *instance or subclass of* instead of the declared *instance of*. The contemporary check then tests whether the lifespans of subject and named entity overlap: the entity’s inception must not fall after the subject’s death, and its dissolution, where it has one, not before the subject’s birth. Entities carrying no inception date (P571) are excluded. The check is applied only to answers that passed the value-type check, since an answer of an impermissible type cannot be meaningfully dated, and testing it would introduce additional error. Both checks read constraint definitions from a snapshot of 14 June 2026, and all SPARQL queries ran against the QLever mirror of Wikidata to avoid the query service’s rate limits [20].

One counting rule within that check materially affects the rates we report. Where an entity carries several inception dates, we take the most recent. Germany (Q183) carries dates from 641 BC to 23 May 1949, so an LLM that assigns Johann Wolfgang von Goethe (Q5879) German citizenship is recorded as a contemporary violation. The model may have intended a predecessor state later absorbed into Germany, in Goethe’s case Saxe-Weimar-Eisenach (Q155570), but the strict framing asks explicitly for a country of citizenship, and the modern name answers that question imprecisely. Section 5 reports the extent to which our headline rate depends on this rule.

4. Evaluation

Three quantities bound all results reported below. We posed 7,264 questions. Of the answers, 6,142 reached the value-type check, and 4,101 of those reached the contemporary check. Writing N for the total number of answers obtained, E_{vt} and E_{ct} for those eligible for the value-type and the contemporary check, and V_{vt} and V_{ct} for the violations identified by each check:

$$\text{coverage} = \frac{E_{vt}}{N}, \quad r_{vt} = \frac{V_{vt}}{E_{vt}}, \quad r_{ct} = \frac{V_{ct}}{E_{ct}}.$$

The first gap, from N to E_{vt} , is coverage: an answer must survive extraction, link to an entity, and carry instance of (P31) metadata before any check can be applied to it. The second gap, from E_{vt} to E_{ct} , is narrower and has three causes. Only four of the six properties carry a contemporary constraint; the check is applied only to answers that already passed the value-type check; and a further 76 answers are excluded because the linked entity has no inception date or the subject lacks usable lifespan dates.

Coverage is an eligibility condition, not a hallucination metric. It indicates what proportion of the experiment is observable, not what proportion is correct, and it bounds every rate that follows: the rates describe the answers that could be linked, rather than the full set of answers the LLMs generated. Confidence intervals, where reported, are 95% Wilson intervals and are descriptive. Each quantifies the sampling uncertainty of one rate in isolation, under an independence assumption that our design only approximates, since all four LLMs answer the same questions and answers about the same subject are not independent. We do not interpret them as significance tests of differences between rates.

4.1. Coverage

1,122 answers never reached the check, 15.4% of the questions posed. Five mechanisms remove an answer from the evaluable set. An answer may be unlinkable: a string is extracted, but no Wikidata entity matches it. The LLM may abstain, stating that it does not know the subject. It may refuse semantically, asserting that the subject has no such property, which differs from abstention and is counted separately. Extraction may fail on prose too long or too weakly formatted to reduce to a single string. A small remainder falls into other categories: a linked entity carrying no instance of (P31) value, a parser that returns the subject’s own name instead of the property value, or, on P27 alone, an answer that names two countries. Unlinkable answers outnumber the other four categories combined,

at 609 cases, and their composition is informative: 318 are compound role or title phrases of the form “Director of the Pasteur Institute”, and a further 163 are short descriptive titles such as “First Painter to the Emperor”. These answers mostly denote real referents and are often correct. Wikidata has a node for the role and a node for the institute, and none for the phrase that joins the two.

Coverage is not distributed evenly across properties or models (Figure 1). Country of citizenship (P27) is the most evaluable property by a wide margin, at 93.3% [91.9–94.4], which is unsurprising: the set of countries is small and closed, so naming one leaves little scope for extraction or linking failure. Position held (P39) is lowest, at 71.8% [68.8–74.6], with most of its unevaluable answers arising from the compound-role problem. The four properties in between form a cluster whose intervals overlap, so we do not infer a clear ordering within it.

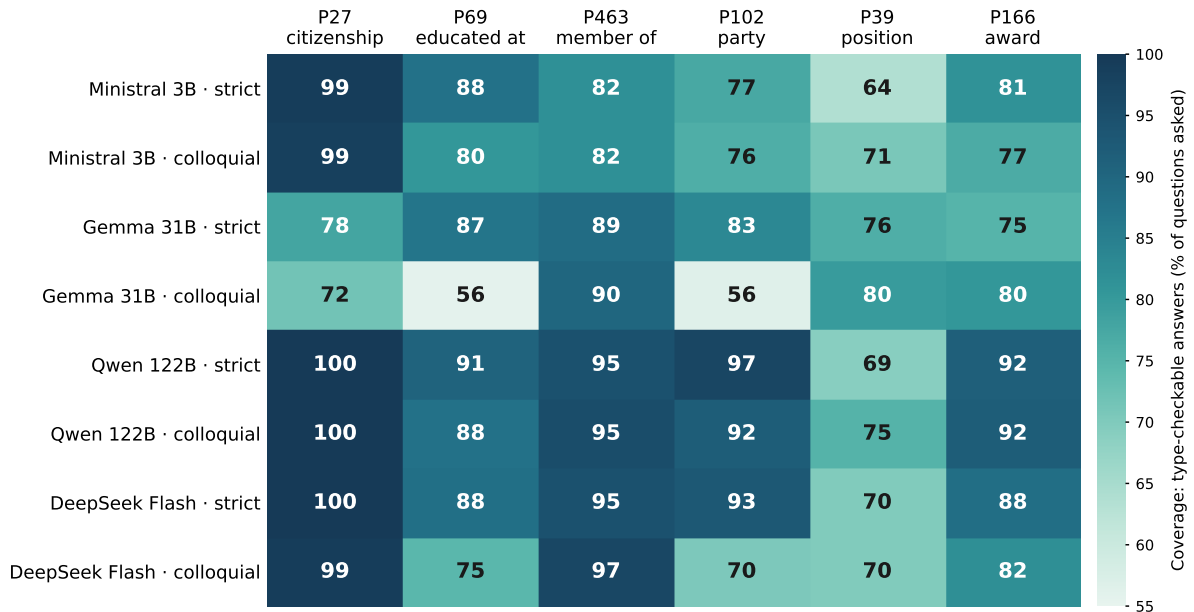


Figure 1: Coverage by LLM, prompt framing, and property. Each cell gives, for one LLM and framing on one property, the percentage of questions asked whose answers reached the constraint check. Higher values mean more answers could be evaluated.

Across models the two larger LLMs achieve the highest and most uniform coverage, Qwen at 91.5% and DeepSeek Flash at 86.9%. The lower end of that ranking requires qualification. Gemma has the lowest coverage, at 76.7%, but 45% of its unevaluable answers are abstentions rather than answers that cannot be linked. It is not the least accurate LLM in the panel but the most cautious, and coverage records only how many answers reach the check.

Framing affects coverage less than either property or model size, and it does so in a specific manner. The strict framing reaches 86.6% [85.5–87.7] against 82.5% [81.2–83.7] for the colloquial framing, a gap that is large relative to the interval widths. The unlinkable rate, however, is identical under both framings, at 8.4%. The looser wording therefore costs answers rather than linkability: abstentions increase, and extraction fails more often on the longer prose it produces. Popularity operates similarly, shifting the failure mode rather than the total. Net coverage is 83.8% for obscure subjects against 85.2% for popular ones, and beneath that near-parity, obscure subjects yield three times as many refusals, 6.9% against 2.2%, while popular ones yield more unlinkable answers, 12.5% against 9.0%.

4.2. Violation rates

Of the 6,142 evaluable answers, 891 violate a constraint, which is 14.5%. That figure is driven largely by a single property. Country of citizenship (P27) alone accounts for 79% of all contemporary violations, and with P27 excluded entirely, the rate falls to 6.0%. The counting rule for P27 is given in Section 3.

Considered as two separate rates, r_{vt} stands at 2.2% [1.8–2.6] across all six properties and r_{ct} at 18.5% [17.3–19.7] across the four that carry the constraint. The underlying counts are 134 and 757, a ratio of more than five to one that overstates the difference. Excluding P27 from both renders the two constraint types more comparable, at 118 value-type against 158 contemporary violations, although they are then counted over different property sets, five for value-type and three for contemporary.

Table 2 gives the rates by property, with the breakdown by prompt framing alongside each. Value-type violations are highest for award received (P166), at 4.6%. Contemporary violations are highest for P27, at 40.1% [37.7–42.6], more than four times the rate of the next property. That rate is stable across the panel: 42.4% for Ministral, 40.4% for Qwen, 39.8% for DeepSeek Flash, 37.3% for Gemma. This is a range of five percentage points, small relative to the interval widths.

Prop.	Label	n	Rate [95% CI]	Strict	Colloq.
<i>Value-type (r_{vt})</i>					
P166	award received	1,262	4.6 [3.6–5.9]	2.2	7.0
P39	position held	649	3.4 [2.2–5.1]	4.4	2.4
P102	political party	652	2.8 [1.8–4.3]	1.1	4.7
P69	educated at	1,101	1.7 [1.1–2.7]	1.0	2.6
P27	citizenship	1,522	1.1 [0.6–1.7]	0.8	1.3
P463	member of	956	0.1 [0.0–0.6]	0.2	0.0
All properties		6,142	2.2 [1.8–2.6]	1.4	3.0
<i>Contemporary (r_{ct})</i>					
P27	citizenship	1,493	40.1 [37.7–42.6]	38.9	41.4
P463	member of	926	7.7 [6.1–9.6]	11.0	4.5
P102	political party	619	5.7 [4.1–7.8]	5.6	5.8
P69	educated at	1,063	4.9 [3.7–6.4]	5.7	3.9
All properties		4,101	18.5 [17.3–19.7]	18.6	18.3

Table 2

Violation rates by property in percent, pooled over the four LLMs, with 95% Wilson confidence intervals and the split by prompt framing.

Prompt framing affects the two constraints differently. From strict to colloquial r_{vt} more than doubles, 1.4% against 3.0%, and the most pronounced single case is P166, where the colloquial “name one honour” yields 7.0% against 2.2% for the strict “name one award”. Over the same shift r_{ct} is essentially unchanged, 18.6% against 18.3%.

4.3. Error mechanisms

A violation rate records how frequently a constraint is violated, but not in what manner. We therefore hand-coded every value-type violation and every contemporary violation outside P27, verifying each against the subject’s recorded values in Wikidata, and classified them into four mechanisms:

1. **Relational fabrication.** An assertion of a false edge to an existing but unconnected entity, such as a nineteenth-century scholar placed at an institute founded in 1969.
2. **Generic fallback.** A hypernym supplied where a specific instance was available, such as “Gymnasium” in place of the named school.
3. **Boundary confusion.** A closely related and often correct entity that falls outside the schema boundary: the modern successor of an institution that existed in the subject’s lifetime, or the academy that confers an award where the award itself was requested.
4. **Pipeline artifact.** An error introduced by the pipeline rather than by the model, either a correct answer attached to the wrong node or a Wikidata inception date that records a reorganisation instead of a founding. One such case is the Freemasons being linked to a band of that name founded in 2005.

Model	Params	n	Fab.	Bound.	Gen.	Pipe.	Pooled
<i>Contemporary, excluding P27</i>							
Ministral	3B	609	9.0	1.6	0.0	0.3	11.0
Gemma	31B	598	3.3	1.2	0.0	1.5	6.0
Qwen	122B	719	0.7	1.8	0.0	1.9	4.5
DeepSeek	284B	682	1.2	1.0	0.0	1.2	3.4
Violations (n)			88	37	0	33	158
<i>Value-type, all six properties</i>							
Ministral	3B	1,509	0.5	0.7	1.2	0.1	2.5
Gemma	31B	1,393	0.0	2.1	1.0	0.0	3.1
Qwen	122B	1,661	0.0	0.7	0.6	0.1	1.3
DeepSeek	284B	1,579	0.0	1.7	0.1	0.3	2.0
Violations (n)			8	77	43	6	134

Table 3

The hand-coded violations decomposed by mechanism and model size, each in percent of that LLM’s answers eligible for the check in question, given by n, with the underlying counts in the last row of each panel. The contemporary panel is pooled over educated at (P69), member of political party (P102) and member of (P463). Fab.: relational fabrication. Bound.: boundary confusion. Gen.: generic fallback. Pipe.: pipeline artifact. Pooled gives the combined rate; components may not sum to it exactly due to rounding.

The two constraints have opposite profiles (Table 3). Among the 158 contemporary violations outside P27, 88 are fabrications against 37 boundary confusions. Among the 134 value-type violations the proportions are inverted, 8 fabrications against 77 boundary confusions. The generic column is empty in the upper panel for a mechanical reason rather than a behavioural one: generic terms rarely carry an inception date, so they are neither dated nor counted.

Read by model size, the mechanisms diverge. Fabrication declines sharply, from 9.0% at 3B to 0.7% and 1.2% in the two largest models. Boundary confusion does not, remaining between 1.0% and 1.8% at every size. Generic fallback decreases on the value-type side, from 1.2% at 3B to 0.1% at 284B. The pipeline artifact rate moves in the opposite direction, increasing with size, 0.3% at 3B against 1.9% at 122B. This does not represent a defect that grows with scale. Larger models name more existing and more specific entities, and a specific entity can be attached by our linker to the wrong node, whereas a vague one offers no such opportunity for error.

Both readings require qualification from the intervals reported in the pooled column. On the value-type side the extremes lie far apart relative to the interval widths, Gemma at 3.1% [2.3–4.1] and Qwen at 1.3% [0.9–2.0], while Ministral and DeepSeek Flash lie close to both of their neighbours. On the contemporary side Ministral at 11.0% [8.8–13.7] lies well above the other three, whose intervals all overlap: Gemma at 6.0% [4.4–8.2], Qwen at 4.5% [3.2–6.2] and DeepSeek Flash at 3.4% [2.3–5.0]. The threefold drop from the smallest model to the largest therefore reflects largely a single contrast rather than a smooth gradient across four sizes.

5. Discussion

Our central result is not only that LLM outputs violate Wikidata constraints, but that the constraints themselves can be repurposed as an evaluation instrument. Most work employs a knowledge graph as a source of facts or context; we instead use its community-maintained rules as an external audit of LLM output. This provides a distinct evaluation signal: whether a generated claim respects the graph’s temporal and ontological boundaries. Once an answer is grounded, the check is deterministic, locally explainable, and requires neither a ground-truth answer, a source document, nor a second LLM as judge.

The 14.5% violation rate reported in Section 4 is not an overall hallucination rate, for two reasons. First, the check is not complete as a test of factual correctness: a claim may be false while still naming an

entity of the permitted type and period, and conversely a constraint flag may reflect a grounding error, incomplete metadata, or a debatable modelling choice rather than a genuine hallucination. Second, the headline figure is strongly influenced by our strict treatment of modern country names for historical persons; excluding country of citizenship (P27) alone, the rate falls to 6.0%. The 14.5% figure should therefore be interpreted as the prevalence of selected constraint violations among evaluable answers, not as a general-purpose factual-error rate.

5.1. What the constraints reveal

The two constraint types expose different forms of imprecision. For the three contemporary properties other than P27, the pooled violation rate falls from 11.0% for Ministral to 3.4% for DeepSeek. The manual decomposition indicates that relational fabrication decreases from 9.0% in the smallest model to approximately 1% in the two largest, whereas temporal boundary confusion remains near 1–2%. The latter typically occurs when a model names the modern successor of an institution that existed in an earlier form. P27 is an extreme instance of the same pattern: all four models frequently return a present-day state where Wikidata records a historical predecessor.

These results should not be interpreted as a strict scaling law, given that parameter counts are confounded by architectural and training differences across the four model families, and overlapping confidence intervals render distinctions among the three larger models statistically marginal. Rather, the central qualitative finding is a clear divergence in failure modes: while scaling substantially reduces relational fabrications, it does not suppress temporal boundary errors. Value-type violations, by contrast, do not order consistently by model size, whereas prompt framing shows a clearer association: strict prompts reduce the pooled rate from 3.0% to 1.4%.

Across both constraints, the recurring failure is boundary confusion: the answer denotes an existing and relevant entity that lies just outside the required temporal or ontological boundary. These plausible boundary errors are of particular interest because they differ from outright invention and are difficult for human readers to detect. The results therefore motivate treating constraint adherence as a separate evaluation dimension alongside factual accuracy and contextual faithfulness.

5.2. Scope and implications

The audit does not constitute a test of truth. A false claim may pass if its object has the correct type and time period, while a valid claim may be flagged owing to an incorrect link, incomplete metadata, or a debatable constraint interpretation. Violations should therefore trigger retrieval, review, or regeneration rather than automatic rejection.

Coverage is also part of the measurement. The audit reached 6,142 of 7,264 answers; unlinkable compound roles and verbose titles accounted for most unevaluable answers. Because rates are conditional on coverage, models that abstain more frequently are evaluated on an easier subset, while detailed answers create more opportunities for both detection and mis-linking. Moreover, 33 of 158 non-P27 contemporary flags and six of 134 value-type flags were pipeline artifacts. Future benchmarks should therefore report coverage, conditional violation rates, and independently annotated flag precision together.

The study is limited to six properties, two constraints, historical humans, and one generation per prompt; P102 uses a separate population, the sample remains male-skewed, and parts of the error analysis required manual judgement. The results establish feasibility and recurring failure patterns, but not a causal effect of model scale or a general estimate of LLM factuality.

6. Conclusion and Future Work

We presented a post-generation audit that grounds LLM answers in Wikidata and checks them against value-type and contemporary constraints. Its central contribution is the repurposing of rules created for KG quality control as an external, transparent standard for generated claims. The experiments indicate

that this signal detects not only fabrications, but also plausible temporal and ontological boundary errors that aggregate factuality measures may obscure.

While this study establishes the feasibility of constraint-based auditing, it is bounded by several limitations. The evaluation is restricted to six properties, two constraint types, and one generation per prompt over a primarily male-skewed sample of historical humans (with P102 evaluated on a separate set). Furthermore, the aggregate violation rate is heavily influenced by our conservative counting rule for modern country names (P27), and the error analysis relies in part on manual judgment. Consequently, the results highlight recurring failure modes rather than providing a comprehensive estimate of general LLM factuality or strict scaling laws.

Future work should improve and independently evaluate grounding, especially for compound roles and historical names, and refine temporal validation using predecessor–successor relations, qualifiers, and alternative inception dates. Broader studies should extend to additional constraints, entity classes, knowledge graphs, repeated generations, and controlled model comparisons. Finally, integrating the audit into inference could allow violations to trigger evidence retrieval, abstention, regeneration, or structured repair.

Acknowledgments

This research was funded in whole or in part by the Austrian Science Fund (FWF) 10.55776/COE12. For open access purposes, the author has applied a CC BY public copyright license to any author accepted manuscript version arising from this submission.

Declaration on Generative AI

During the preparation of this work, the authors used Claude for grammar and spelling checking and for paraphrasing and rewording, and Claude Code for code generation. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

References

- [1] A. Hogan, E. Blomqvist, M. Cochez, C. D’amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A.-C. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, A. Zimmermann, Knowledge graphs, *ACM Computing Surveys* 54 (2021) 1–37. URL: <http://dx.doi.org/10.1145/3447772>. doi:10.1145/3447772.
- [2] A. T. Kalai, O. Nachum, S. S. Vempala, E. Zhang, Why language models hallucinate, 2025. URL: <https://arxiv.org/abs/2509.04664>. arXiv:2509.04664.
- [3] A. Singla, A. Sukharevsky, B. Hall, L. Yee, M. Chui, T. Balakrishnan, The state of AI in 2025: Agents, innovation, and transformation, 2025. URL: <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai>.
- [4] W. Wang, Z. Ma, M. Ding, S. Zheng, S. Liu, J. Liu, J. Ji, W. Chen, X. Li, L. Shen, Y. Yuan, Medical reasoning in the era of LLMs: A systematic review of enhancement techniques and applications, 2025. URL: <https://arxiv.org/abs/2508.00669>. arXiv:2508.00669.
- [5] A. Bott, A. Wong, Evolution of Power and Energy Insurance Claims Handling in the Age of Artificial Intelligence, Technical Report, Wotton Kearney and Chubb Insurance (Australia), 2025. URL: <https://www.wottonkearney.com/wp-content/uploads/2026/03/Evolution-of-Power-and-Energy-Insurance-Claims-Handling-in-the-Age-of-Artificial-Intelligence.pdf>.
- [6] N. Ferranti, J. F. D. Souza, S. Ahmetaj, A. Polleres, Formalizing and validating Wikidata’s property constraints using SHACL and SPARQL, *Semantic Web* 15 (2024) 2333–2380. doi:10.3233/SW-243611.

- [7] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, X. Wu, Unifying large language models and knowledge graphs: A roadmap, *IEEE Transactions on Knowledge and Data Engineering* 36 (2024) 3580–3599. doi:10.1109/TKDE.2024.3352100.
- [8] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. Ness, J. Larson, From local to global: A Graph RAG approach to query-focused summarization, 2025. URL: <https://arxiv.org/abs/2404.16130>. arXiv:2404.16130.
- [9] L. Pusch, T. O. F. Conrad, Combining LLMs and knowledge graphs to reduce hallucinations in question answering, 2025. URL: <https://arxiv.org/abs/2409.04181>. arXiv:2409.04181.
- [10] F. Darari, J. G. Ongiris, B. C. L. Tobing, D. R. Faisal, O. Lee, Leveraging knowledge graphs to mitigate counting hallucinations: A case study of Wikidata, in: *Proceedings of the 5th Wikidata Workshop (Wikidata 2025) co-located with the 24th International Semantic Web Conference (ISWC 2025)*, volume 4108 of *CEUR Workshop Proceedings*, CEUR-WS.org, Nara, Japan, 2025. URL: <https://ceur-ws.org/Vol-4108/paper1.pdf>.
- [11] L. Luo, Z. Zhao, G. Haffari, Y.-F. Li, C. Gong, S. Pan, Graph-constrained reasoning: Faithful reasoning on knowledge graphs with large language models, 2025. URL: <https://arxiv.org/abs/2410.13080>. arXiv:2410.13080, accepted at ICML 2025.
- [12] G. Agrawal, T. Kumara, Z. Alghamdi, H. Liu, Can knowledge graphs reduce hallucinations in LLMs? : A survey, 2024. URL: <https://arxiv.org/abs/2311.07914>. arXiv:2311.07914, accepted at NAACL 2024.
- [13] A. Robertson, H. Liang, M. Gani, R. Kumar, S. Rajamohan, KGHaluBench: A knowledge graph-based hallucination benchmark for evaluating the breadth and depth of LLM knowledge, 2026. URL: <https://arxiv.org/abs/2602.19643>. arXiv:2602.19643, also published in *Findings of the Association for Computational Linguistics: EACL 2026*.
- [14] H. Sansford, N. Richardson, H. P. Maretic, J. N. Saada, GraphEval: A knowledge-graph based LLM hallucination evaluation framework, 2024. URL: <https://arxiv.org/abs/2407.10793>. arXiv:2407.10793.
- [15] A. W. Lee, J. Chan, M. Fu, N. Kim, A. Mehta, D. Raghavan, U. Çetintemel, Semantic integrity constraints: Declarative guardrails for AI-augmented data processing systems, *Proceedings of the VLDB Endowment* 18 (2025) 4073–4080. URL: <http://dx.doi.org/10.14778/3749646.3749677>. doi:10.14778/3749646.3749677.
- [16] Wikidata contributors, value-type constraint (q21510865) — Wikidata, 2026. URL: <https://www.wikidata.org/wiki/Q21510865>.
- [17] Wikidata Contributors, Help:property constraints portal/Contemporary, 2021. URL: https://www.wikidata.org/wiki/Help:Property_constraints_portal/Contemporary.
- [18] K. J. Preacher, D. D. Rucker, R. C. MacCallum, W. A. Nicewander, Use of the extreme groups approach: A critical reexamination and new recommendations, *Psychological Methods* 10 (2005) 178–192. doi:10.1037/1082-989X.10.2.178.
- [19] M. Sclar, Y. Choi, Y. Tsvetkov, A. Suhr, Quantifying language models’ sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting, in: *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL: <https://arxiv.org/abs/2310.11324>.
- [20] H. Bast, B. Buchhold, QLever: A query engine for efficient SPARQL+Text search, in: *Proceedings of the 2017 ACM Conference on Information and Knowledge Management (CIKM)*, ACM, Singapore, 2017, pp. 647–656. doi:10.1145/3132847.3132921.

A. Online Resources

The implementation and the evaluation data are available via

- GitHub repository: <https://github.com/eddievanhalensolo/kg-halu-audit>