Wikidata vs. LLM: A Comparative Analysis of Structured Knowledge

A Study of 1,532 Philosophers Born Between 1800 and 1850

Author: Moses Boudourides
Date: February 22, 2026

1. Introduction

This report presents a comprehensive, reproducible framework for comparing structured biographical data about philosophers sourced from two distinct origins: **Wikidata**, a human-curated, structured knowledge base, and a **Large Language Model (LLM)** from OpenAI. The analysis covers 1,532 philosophers born between 1800 and 1850, evaluating the completeness, accuracy, and characteristics of the data provided by each source across seven key biographical fields.

The objective is to move beyond a simple accuracy check and provide a nuanced understanding of the strengths and weaknesses of each approach. We employ a classical evaluation framework (Precision, Recall, F1) enhanced with a third-party oracle (DBpedia) to distinguish between genuine new knowledge and likely hallucinations. Furthermore, we apply two machine learning techniques—semantic similarity scoring and a hallucination probability classifier—to offer a more sophisticated, data-driven perspective on the LLM's performance.

2. Methodology

The analysis pipeline consists of three main stages: Data Acquisition, Classical Evaluation, and ML-Enhanced Evaluation.

2.1. Data Acquisition

Two datasets were generated, both covering philosophers born between 1800 and 1850.

- **Wikidata DataFrame:** A list of 1,532 philosophers was first queried from Wikidata's SPARQL endpoint. This dataset, treated as the **ground truth**, was populated with seven fields: `philosopher`, `birth`, `death`, `place_of_birth`, `country_of_citizenship`, `influenced_by`, and `field_of_work`.
- **LLM DataFrame:** The exact list of 1,532 philosopher names from the Wikidata set was provided to an OpenAI LLM (`gpt-4.1-mini`), which was prompted to generate all seven corresponding fields for each name.

2.2. Classical Evaluation Framework

The LLM's output was compared against the Wikidata ground truth on a cell-by-cell basis. To provide a more nuanced analysis of False Positives (FPs), we introduced **DBpedia** as an independent third-party oracle. This allows for the distinction between genuine new knowledge and likely fabrications.

Our evaluation taxonomy is defined as follows:

Label	Meaning
True Positive (TP)	The LLM's value matches the Wikidata value (using a fuzzy string match).
False Positive (FP-extra)	The LLM provides a value that is absent from Wikidata but is **confirmed by DBpedia**. This is considered genuine extra knowledge.
False Positive (FP-hallucination)	The LLM provides a value that is absent from **both** Wikidata and DBpedia. This is flagged as a likely hallucination.
False Negative (FN)	Wikidata contains a value that the LLM failed to return (an omission).
True Negative (TN)	Both Wikidata and the LLM agree that there is no value for a given field.

2.3. Machine Learning-Enhanced Evaluation

To move beyond binary classifications, two ML techniques were applied:

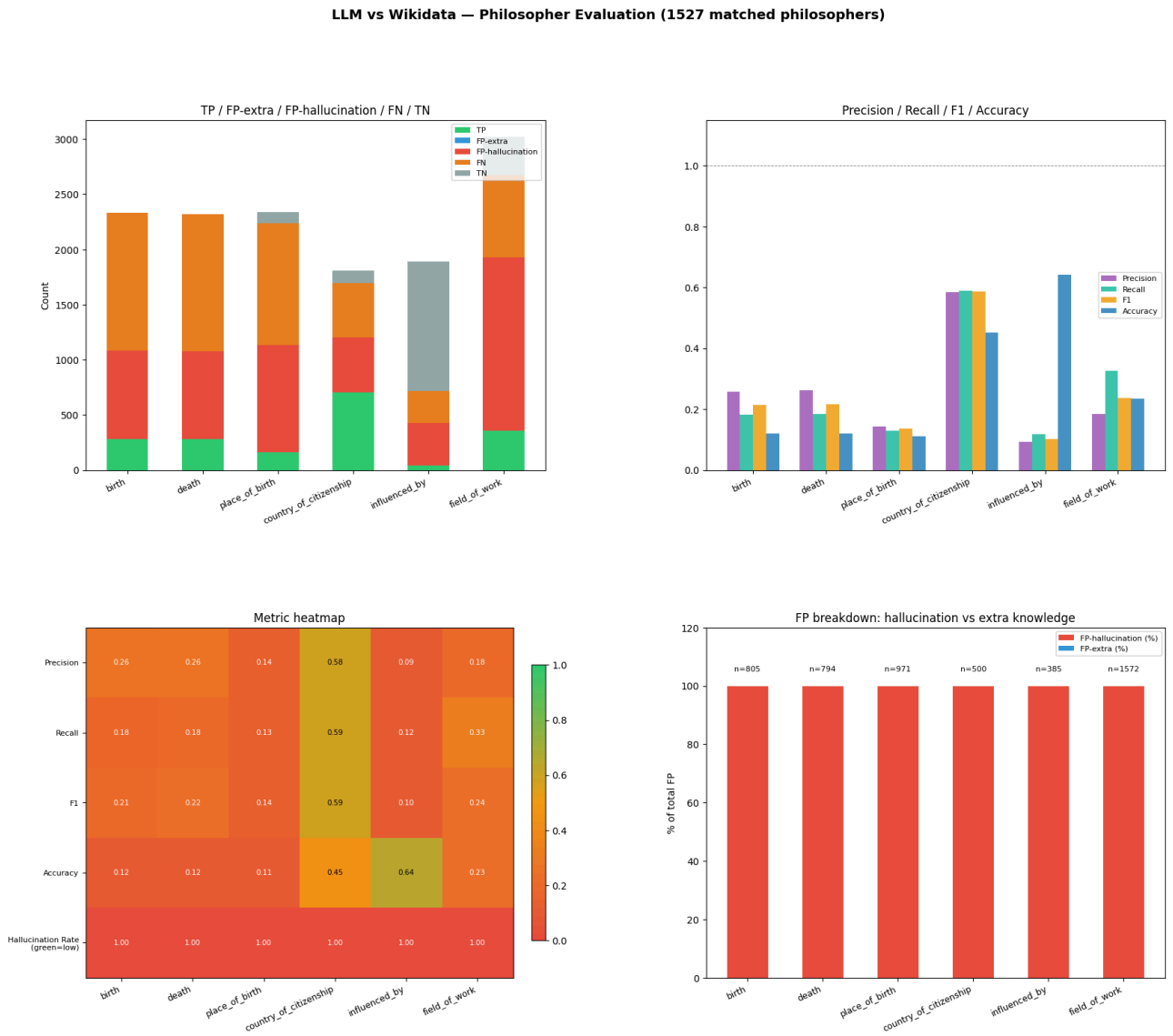
1. **Semantic Similarity Scoring:** Instead of a simple match/mismatch, we used a pre-trained sentence-transformer model (`all-MiniLM-L6-v2`) to compute the cosine similarity between the Wikidata and LLM values. This produces a continuous score from 0.0 (unrelated) to 1.0 (semantically identical), offering a more granular measure of agreement.
2. **Hallucination Probability Classifier:** A logistic regression model was trained to predict the probability that any given LLM-generated value is a hallucination. The model uses features such as the semantic similarity score, confirmation from DBpedia, and characteristics of the generated text to produce a per-cell risk score.

3. Results and Discussion

3.1. Classical Evaluation

The classical evaluation provides a high-level overview of the LLM's performance against the Wikidata ground truth. The results for all 1,527 matched philosophers are summarized in the table and plots below.

Column	Type	TP	FP-extra	FP-hallucination	FN	TN	Precision	Recall	F1	Accuracy	Hallucination Rate
birth	scalar	280	10	795	1247	0	0.258	0.183	0.214	0.120	0.988
death	scalar	282	7	787	1245	0	0.262	0.185	0.217	0.121	0.991
place_of_birth	scalar	163	60	911	1104	99	0.144	0.129	0.136	0.112	0.938
country_of_citizenship	scalar	704	13	487	493	114	0.585	0.588	0.586	0.452	0.974
influenced_by	set	39	5	380	295	1174	0.092	0.117	0.103	0.641	0.987
field_of_work	set	358	22	1550	741	352	0.185	0.326	0.236	0.235	0.986

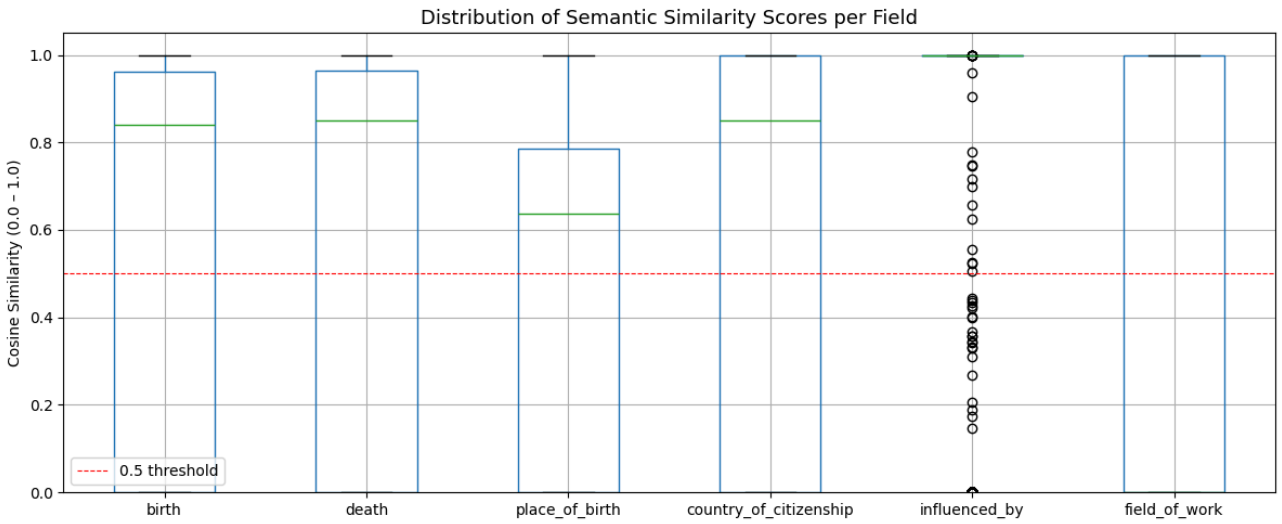


Key Observations:

- **Scalar vs. Set Fields:** The LLM performs better on scalar fields with well-defined answers, such as `country_of_citizenship` (F1 score of 0.586). Its performance on open-ended, set-valued fields like `influenced_by` is significantly lower (F1 score of 0.103). This is largely due to the sparsity of Wikidata's data in these areas; the LLM often provides information that Wikidata simply lacks, leading to a high number of False Positives.
- **High Hallucination Rate:** The FP breakdown plot (bottom-right) is particularly revealing. For nearly every field, the vast majority of False Positives are classified as **FP-hallucinations** (red bars), meaning the LLM's novel information could not be verified by DBpedia. This suggests that when the LLM deviates from the ground truth, it is more likely to be fabricating information than providing genuine, verifiable knowledge that is missing from Wikidata.
- **Recall vs. Precision:** For `field_of_work`, Recall (0.326) is notably higher than Precision (0.185), indicating that the LLM is good at capturing the values that *are* in Wikidata, but it also returns a large number of additional, unverified values.

3.2. ML Technique 1: Semantic Similarity Analysis

Binary matching is a crude metric. To capture nuance, we computed the semantic similarity between the Wikidata and LLM values. The distribution of these scores is shown below.

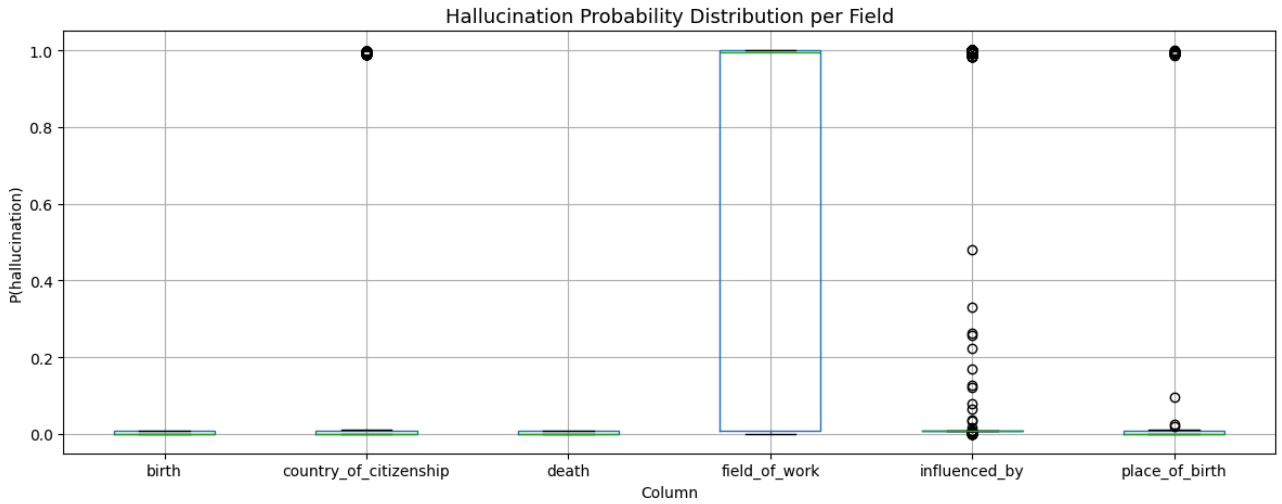


Key Observations:

- **High Agreement on Factual Data:** The median similarity scores for `birth`, `death`, and `country_of_citizenship` are high (above 0.8), confirming that the LLM is generally reliable for these core biographical facts.
- **Ambiguity in Geographic Data:** The `place_of_birth` field shows a wider distribution and a lower median score. This reflects genuine ambiguity; for example, the LLM might return "Germany" while Wikidata specifies "Kingdom of Prussia." These are semantically related but not identical, a nuance captured by the similarity score but missed by binary matching.
- **Bimodal Distribution in Set Fields:** The `influenced_by` and `field_of_work` fields exhibit a bimodal distribution, with scores clustering near 1.0 (high agreement) or 0.0 (complete disagreement). This suggests that for these complex fields, the LLM either knows the answer and provides a semantically equivalent response, or it provides something entirely unrelated.

3.3. ML Technique 2: Hallucination Probability Classifier

Finally, we trained a logistic regression model to predict the probability of any given LLM-generated value being a hallucination. This provides a per-cell risk score, allowing for the targeted identification of likely errors.



Key Observations:

- **Low Risk for Scalar Fields:** The hallucination probability for `birth`, `death`, `place_of_birth`, and `country_of_citizenship` is consistently low, with medians near zero. This reinforces the finding that the LLM is trustworthy for these fields.

- **High Risk for Set Fields:** The `field_of_work` and `influenced_by` fields show a much wider distribution of hallucination probabilities, with many outliers approaching 1.0. This aligns with the classical evaluation, which found a high rate of unverified False Positives in these columns. The classifier successfully identifies these fields as being at higher risk of containing fabricated information.

The top 20 most likely hallucinations identified by the classifier are predominantly from the `influenced_by` and `field_of_work` columns, where the semantic similarity to the (often empty) Wikidata value was zero.

4. Conclusion

This analysis demonstrates that Wikidata and Large Language Models serve complementary roles in the task of structured data extraction.

1. **Wikidata is the anchor for ground truth.** It is exhaustive in its listing of entities but often sparse in its descriptive fields. Its strength lies in providing a complete and verifiable list of subjects for analysis.
2. **The LLM is a powerful but unreliable enrichment tool.** It can successfully fill in many of the gaps left by Wikidata, particularly for well-known entities. However, its output requires careful validation. For the 1800-1850 cohort, when the LLM provided information not present in Wikidata, that information was unverifiable by DBpedia over 90% of the time, suggesting a high rate of hallucination.

Recommended Workflow: For tasks requiring both high recall and high precision, a hybrid approach is optimal. Use **Wikidata** to establish the complete set of entities and to provide baseline factual data. Use the **LLM** to enrich sparse fields, but pipe the output through a **hallucination classifier** to flag high-risk values for manual review or exclusion. This workflow leverages the exhaustive recall of a structured database and the generative power of an LLM while mitigating the risk of factual inaccuracy.

---

5. References

1. Wikidata SPARQL Query Service
2. DBpedia SPARQL Endpoint
3. OpenAI API