Ontology-Grounded Language Modeling: Enhancing GPT-Based Philosophical Text Generation with Structured Knowledge

Claire Ponciano[a], Markus Schaffert[b] and Jean-Jacques Ponciano[c]
i3mainz, University of Applied Sciences, Germany

Keywords: Ontology-Grounded Language Modeling, GPT, Knowledge-Enhanced Text Generation, Retrieval-Augmented Generation, Spinoza, Linked Open Data, Historical Text Synthesis, Philosophical Language Modeling, BERTScore Evaluation, Structured Knowledge Integration, Latin Text Generation, Large Language Models, Text Style Transfer, Semantic Conditioning, Canonical Corpus Fine-Tuning.

Abstract: We present an ontology-grounded approach to GPT-based text generation aimed at improving factual grounding, historical plausibility, and stylistic fidelity in a case study: Baruch Spinoza's Latin writings. We construct a compact ontology from Linked Open Data (Wikidata/DBpedia) augmented with expert-curated facts, serialize triples into natural-language statements, and interleave these with a canonical Latin corpus during fine-tuning of a GPT-2 (124M) model. At inference, retrieval-augmented generation (RAG) prepends ontology-derived facts and lightweight stylistic instructions, guiding the model toward historically consistent continuations in Spinoza's register. Evaluation follows an 80/20 paragraph split of Ethica: we generate continuations for the 80% of segments retained and measure the semantic similarity (BERTScore) with the 20% omitted. This evaluation is completed by an expert assessment of historical plausibility and cosine similarity scores computation for the stylistic authenticity. Relative to a GPT-2 baseline trained only on the Latin corpus, our ontology-grounded variant achieves higher BERTScore and produces fewer factual and conceptual errors, preserving Latin rhetorical structure. These results indicate that structured knowledge integration is a feasible and effective way to make generative models more reliable for cultural-heritage text.

1 INTRODUCTION

The preservation of cultural-heritage texts is hampered by losses due to deterioration and historical events, and by restoration workflows that rely on expert inference, cross-referencing, and fragment interpretation—processes that are time-intensive and subjective. Recent advances in NLP and large language models (LLMs) offer automation, but lack the semantic precision needed to reproduce intricate philosophical and scientific texts. Ontological knowledge bases provide the required structure and contextual grounding.

We propose integrating dynamic ontology generation with LLMs, building on our ODKAR framework (Ontology-Based Dynamic Knowledge Acquisition and Automated Reasoning)(Prudhomme et al., 2024), which uses NLP, OWL, and SWRL to construct ontologies from text. ODKAR-derived triples are serialized as natural-language statements and supplied to the LLM to guide reconstruction. Using Spinoza as a case study, we target *plausible, style-consistent reconstructions of missing passages* rather than literal recovery of lost works.

Our goals are historical authenticity, semantic consistency, and linguistic-philosophical coherence. We evaluate on a comprehensive corpus with approximately 30% held out, asking the system to reconstruct withheld segments under predefined structures. Results quantify restoration accuracy and qualitative coherence, and show that ontology-grounded generation: (i) combines structured semantics with generative modeling for historical restoration, (ii) maintains semantic fidelity and logical consistency via automated processing, and (iii) yields a robust, reproducible framework applicable beyond this case study.

[a] https://orcid.org/0000-0001-8883-8454
[b] https://orcid.org/0000-0002-7970-9164
[c] https://orcid.org/0000-0001-8950-5723

Ponciano, C., Schaffert, M. and Ponciano, J.-J.
Ontology-Grounded Language Modeling: Enhancing GPT-Based Philosophical Text Generation with Structured Knowledge.
DOI: 10.5220/0013864400003985

In *Proceedings of the 21st International Conference on Web Information Systems and Technologies (WEBIST 2025)*, pages 459-467
ISBN: 978-989-758-772-6; ISSN: 2184-3252

2 RELATED WORK

2.1 Personality-Aware Text Generation

Persona-grounded generation conditions models on profile traits, starting with Persona-Chat (Zhang et al., 2018). Transformer baselines (e.g., GPT-2) and adapters such as PsychAdapter (Liu et al., 2023) inject continuous trait embeddings (e.g., Big Five) to yield stylistic consistency (Zheng et al., 2023). Efficient control uses Contrastive Activation Steering and LoRA for style adaptation without full retraining (Zheng et al., 2023; Hu et al., 2021). Evaluation typically combines automatic metrics (BLEU, ROUGE) with human judgments for persona alignment and coherence (Papineni et al., 2002; Lin, 2004; Zhang et al., 2018).

2.2 Multilingual and Low-Resource Persona Modeling

Work has focused largely on English; XPersona broadened coverage and showed the promise of multilingual transformers (Lin et al., 2020). Zero-shot cross-lingual transfer remains difficult due to cultural/linguistic variation (Majumder et al., 2020; Lin et al., 2021; Zheng et al., 2021). For low-resource settings, researchers rely on machine translation, multilingual pretraining (e.g., mT5, XLM-R), and careful fine-tuning or prompting, though methods tailored specifically to sparse supervision are still limited (Lin et al., 2020; Majumder et al., 2020; Hedderich et al., 2021).

2.3 Ontology and Linked Open Data (LOD) for Knowledge-Aware Generation

Ontologies and LOD enable structured data-to-text with semantic rigor (Gardent et al., 2017; Shimorina and Gardent, 2019). Knowledge-graph sources (DBpedia, Wikidata) guide neural generators toward factual fidelity and coverage (Gardent et al., 2017; Ferreira et al., 2020). Transformer-era systems achieve strong accuracy and completeness on WebNLG-style benchmarks, highlighting the value of explicit structure for generation (Gardent et al., 2017; Shimorina and Gardent, 2019).

2.4 Integrating Ontologies with LLMs

LLMs are fluent but prone to hallucinations (Ji et al., 2023). Integrations that surface structured knowledge—e.g., knowledge-enhanced prompting (KELP) and historically informed models (Kongzi)—improve factuality, semantic consistency, and contextual adequacy (Liu et al., 2024; Yao et al., 2023). Multilingual LOD further supplies language-agnostic context that benefits low-resource scenarios (Gardent et al., 2017; Ferreira et al., 2020). Overall, combining ontological structure with LLMs is a promising route to reliable, context-sensitive generation in cultural-heritage applications.

3 METHODOLOGY: ONTOLOGY-INTEGRATED LLM PIPELINE

3.1 Ontology Construction

We first constructed a structured ontology of Baruch Spinoza's life, works, and intellectual milieu leveraging Linked Open Data (LOD) resources such as DBpedia(Auer et al., 2007) and Wikidata (Vrandečić and Krötzsch, 2014). These resources were augmented with manually curated historical facts to fill critical gaps (e.g., Spinoza's Portuguese-Jewish ancestry, his emigration to Amsterdam due to religious persecution, and his excommunication from the Jewish community in 1656). The resulting knowledge was formalized into RDF/OWL triples (McGuinness and Van Harmelen, 2004), capturing semantic relationships such as *influencedBy* (Descartes), *hasEthnicBackground* (Portuguese-Jewish), and *authoredWork* (Ethica). This structured representation facilitated precise semantic querying and integration with the language model.

3.2 Ontology-Grounded Pretraining and Fine-Tuning

We ground the GPT-based model in structured knowledge by converting ontology triples into textual statements and integrating them directly into the fine-tuning corpus (Logan IV et al., 2019; Liu et al., 2021).

Triple-to-Text Conversion Strategy. A lightweight rule-based pipeline maps RDF triples $(subject, predicate, object)$ to grammatical English:

- **Predicate Splitting:** split camel/PascalCase predicates (e.g., `influencedBy` $\rightarrow$ "influenced by"; `excommunicatedOn` $\rightarrow$ "excommunicated on") (Binkley et al., 2009; Allamanis et al., 2021).
- **Template Construction:**

  - if the predicate is verbal, concatenate subject + (auxiliary) + predicate + object.
    Example: (*Ethica*, *authoredBy*, *Spinoza*) → *"The Ethica was authored by Spinoza."*
  - if the predicate is adjectival/nominal, use attributive/possessive verbs (*has*/*had*/*is*).
    Example: (*Spinoza*, *ethnicBackground*, *Portuguese-Jewish*) → *"Spinoza had a Portuguese-Jewish ethnic background."*
- **Named Entities:** preserve canonical capitalization and formatting (Spinoza, *Ethica*, Descartes).

This yields scalable, interpretable factual sentences while preserving ontology semantics.

Corpus Integration and Dataset Preparation. For each training batch, ontology sentences are randomly interleaved with authentic Latin passages so the model jointly learns factual structure and Spinoza's style. The corpus thus contains:

1. **Original Latin:** *"Per Deum intelligo Ens absolute infinitum, hoc est, substantiam constantem infinitis attributis."*
2. **Ontology-Grounded Fact:** "Spinoza originally published many of his works posthumously to avoid religious persecution."

This explicit structuring reduces ambiguity, enforces consistency, and encourages the model to internalize historical relations rather than infer them implicitly.

Model Fine-Tuning Procedure. We fine-tune GPT-2 small (124M) with AdamW (Loshchilov and Hutter, 2019); learning rate 5×10^{-5} (linear schedule, 10% warm-up), batch size 8, for 5–10 epochs. Early stopping monitors validation perplexity on a 10% held-out subset of the Latin corpus to maintain stylistic coherence while injecting factual grounding (Liu et al., 2021).

Example Training Instance. Input: "Spinoza had a Portuguese-Jewish ethnic background. He was excommunicated on 1656. He was influenced by Descartes. *Ethica* was authored by Spinoza. He published posthumously."
Target (Continuation from Original Corpus): *"Per Deum intelligo Ens absolute infinitum, hoc est, substantiam constantem infinitis attributis."*

The GPT-2 model is autoregressive and takes a single concatenated sequence; "input/target" above illustrates the intended continuation context rather than separate encoder/decoder inputs.

3.3 Training Data Clarification

Our fine-tuning corpus comprises (i) Spinoza's Latin texts (*Ethica*, *TTP*, selected letters) and (ii) ontology-derived triple-to-text sentences generated from our knowledge graph. **No additional external prose corpora** were used. The ontology sentences expose factual relations (e.g., *influencedBy*, *authoredBy*) explicitly; the Latin corpus imparts style and rhetoric.

3.4 Ontology-Conditioned Inference

During inference, the GPT-based model leveraged retrieval-augmented generation (RAG) techniques (Lewis et al., 2020) to dynamically condition generated outputs on pertinent ontological knowledge. This process ensured that the model's generated text maintained historical accuracy and factual grounding by explicitly referencing contextually relevant knowledge stored within the ontology.

Dynamic Ontology Retrieval: Given an initial textual prompt provided by the user or an application context, the inference procedure began by querying the ontology dynamically. These queries were executed using standard semantic web querying protocols (e.g., SPARQL) or embedding-based semantic retrieval methods. For example, to generate text "under persecution shortly before Spinoza's death," the system performed the following SPARQL query to retrieve relevant historical facts:

```
PREFIX onto:
SELECT ?event ?date ?detail WHERE {
 ?event onto:concernsPerson onto:Spinoza .
 ?event onto:occurredOnDate ?date .
 ?event onto:hasDetail ?detail .
 FILTER(?date >= "1656"^^xsd:gYear
    && ?date <= "1677"^^xsd:gYear)
 FILTER regex(?detail,
 "persecution|excommunication|censorship",
 "i")}
```

This retrieval resulted in triples such as:

- (*Spinoza*, *excommunicatedOn*, *1656*)
- (*Spinoza*, *publishedPosthumously*, *true*)
- (*Ethica*, *originalLanguage*, *Latin*)
- (*Spinoza*, *influencedBy*, *Descartes*)

Embedding-based retrieval methods alternatively allowed querying via vector similarity, especially useful when handling natural language prompts. For example, embedding the query "Spinoza persecution and death" allowed rapid semantic retrieval of related facts without explicit SPARQL syntax, facilitating more flexible retrieval scenarios.

Prompt Construction with Retrieved Facts: Once retrieved, the ontological facts were synthesized into a structured natural-language context that was prepended directly to the inference prompt provided to the GPT model. This contextual prompt explicitly informed the model of essential historical details, guiding the subsequent text generation. An explicit example of prompt construction from the above retrieved facts is as follows:

> **Constructed Prompt (Contextual Introduction):**
> "Baruch Spinoza was excommunicated by the Jewish community in Amsterdam in 1656 due to his radical philosophical views. Due to fear of religious persecution, he published many of his writings posthumously, including the *Ethica*, originally composed in Latin. Deeply influenced by Descartes, Spinoza further extended rationalist philosophy. The following text, written shortly before his death under persecution, reflects his philosophical reasoning and stylistic approach:"

This detailed contextualization significantly enhanced the generated text's fidelity to Spinoza's historical situation and philosophical lineage.

Prompt Engineering for Stylistic Alignment: Beyond factual grounding, explicit instructions were included to encourage the model to mimic Spinoza's distinctive philosophical and rhetorical style. These prompt engineering techniques were critical in conditioning the model's generative process. For example, explicit stylistic directions embedded within the inference prompt included:

> "The following text should emulate the philosophical argumentation style of Baruch Spinoza, characterized by structured logical reasoning, extensive use of Latin philosophical terminology, and geometric method presentation."

In this study, prompts encouraging stylistic alignment were manually crafted based on domain expertise. However, prompts of comparable effectiveness can also be generated automatically using retrieval-augmented methods or embedding-based similarity techniques. Specifically, by encoding known samples of Spinoza's writing style into vector embeddings, automatic retrieval can identify representative stylistic patterns. These identified patterns can then form the basis of automatically generated prompts that instruct the language model to produce outputs closely aligned with Spinoza's original rhetoric and philosophical methodology. Such automation potentially enhances scalability, reduces manual effort, and ensures consistency across numerous inference tasks.

Such explicit stylistic instructions, coupled with factual grounding provided by retrieved ontology triples, ensured both the historical accuracy and linguistic authenticity of generated texts.

Example of Final Inference Prompt: A comprehensive inference prompt incorporating both factual context and stylistic instruction is exemplified below:

> **Final Prompt Provided to the GPT Model:**
>
> "Baruch Spinoza was excommunicated by the Jewish community in Amsterdam in 1656 due to his radical philosophical views. Fearing religious persecution, he chose to publish many works posthumously, including the *Ethica*, originally composed in Latin. Deeply influenced by Descartes, Spinoza extended rationalist thought significantly beyond his predecessor's bounds. The following Latin text, composed shortly before his death under persecution, must demonstrate Spinoza's philosophical reasoning, structured logical argumentation, and characteristic Latin rhetorical style:"
>
> *[The model-generated Latin philosophical text follows here.]*

This carefully structured prompt ensured the language model's response adhered strictly to historical events, intellectual contexts, and stylistic expectations.

Generation Procedure and Model Parameters: The GPT model generated text using nucleus (top-p) sampling (Holtzman et al., 2019), with $p = 0.9$, ensuring a balance between textual coherence and lexical diversity. We set the maximum generation length to 256 tokens, effectively constraining the model to produce concise, historically plausible narratives without deviation or content drift. Additionally, repetition penalties and controlled decoding methods were used to avoid redundant phrasing and enforce linguistic variability consistent with Spinoza's authentic works.

Through this ontology-conditioned inference process, the language model reliably produced historically coherent and stylistically accurate outputs. Such grounding methodology effectively mitigated common generative model issues like hallucinations and factual inaccuracies, ensuring each generated piece maintained high scholarly integrity and consistency with known historical data.

3.5 Canonical Corpus Fine-Tuning

In parallel with ontology grounding, we fine-tune the model on a curated corpus of Spinoza's authentic Latin to reinforce rhetoric and argument structure. The corpus covers *Ethica*, *Tractatus Theologico-Politicus* (TTP), and selected letters, capturing his geometric method and epistolary register.

Corpus Collection and Selection. Texts were drawn from reliable repositories (Project Gutenberg, Wikisource) and scholarly digitizations to ensure fidelity. Composition:

- ***Ethica, ordine geometrico demonstrata* (1677):** complete treatise with axioms, propositions, corollaries.
- ***Tractatus Theologico-Politicus* (1670):** sustained theological–political argumentation.
- **Letters (1661–1676):** selections exhibiting stylistic and rhetorical variation.

Text Preprocessing and Normalization. We (i) remove marginalia/OCR artifacts; (ii) minimally normalize 17$^{\text{th}}$-century orthography (e.g., *ciuitas* $\rightarrow$ *civitas*, *vnus* $\rightarrow$ *unus*); (iii) segment into sentences/propositions. Example segmentation:

> Original: *"Per Deum intelligo Ens absolute infinitum, hoc est substantiam constantem infinitis attributis. Unumquodque attributum exprimit certam infinitam essentiam aeternam."*
> Segments: (1) *"Per Deum intelligo Ens absolute infinitum, hoc est substantiam constantem infinitis attributis."*
> (2) *"Unumquodque attributum exprimit certam infinitam essentiam aeternam."*

Tokenization Using Byte-Pair Encoding (BPE). A BPE tokenizer (Sennrich et al., 2016) trained on the Latin corpus captures morphological regularities typical of philosophical Latin. Example:

> *"substantiam constantem infinitis attributis"* $\rightarrow$ [*substant*, *iam*, *constant*, *em*, *infinit*, *is*, *attribut*, *is*]

Fine-Tuning Procedure and Hyperparameters. We fine-tune GPT-2 small (124M) with AdamW (Loshchilov and Hutter, 2019); LR 3×10^{-5}, weight decay 0.01, batch size 8, for 5–10 epochs, using early stopping on validation perplexity (10% held-out). This stabilizes convergence on a relatively small corpus while preserving stylistic coherence.

Illustrative Training Example: An explicit example of a fine-tuning training instance is illustrated below:

> **Input (prompt):**
> *"Per Deum intelligo Ens absolute infinitum,"*
>
> **Target (continuation):**
> *"hoc est substantiam constantem infinitis attributis, quorum unumquodque aeternam et infinitam essentiam exprimit."*

This explicit input-target training format enabled the GPT model to learn detailed continuations characteristic of Spinoza's logical argumentation structure, linguistic style, and specific vocabulary.

Outcome and Intended Effect. Stylistic fine-tuning consolidates Spinoza's Latin (vocabulary, syntax, geometric exposition). Combined with ontology-grounded facts, the model produces historically grounded, stylistically faithful generations closely aligned with known texts.

3.6 Text Generation Evaluation

To assess the performance and efficacy of our ontology-grounded GPT model, we conducted an extensive evaluation across three distinct but complementary dimensions: *stylistic alignment*, *historical plausibility*, and *factual grounding*. Each dimension utilized specific methods, metrics, and expert validation processes to ensure comprehensive coverage of evaluation criteria.

1. Stylistic Alignment: Stylistic alignment assessed how closely generated texts conformed to Spinoza's authentic linguistic and rhetorical style. To quantify stylistic similarity objectively, we employed sentence-level embedding similarity metrics using pretrained multilingual language models (e.g., multilingual Sentence-BERT) (Reimers and Gurevych, 2019). Embeddings of generated texts were compared against embeddings from authentic Spinoza texts to calculate cosine similarity scores. For instance:

> Generated Latin text:
> *"Ens infinitum absolute intellegi debet, cuius substantia infinitis attributis exprimitur..."*
>
> Original Spinoza text:
> *"Per Deum intelligo Ens absolute infinitum, hoc est substantiam constantem infinitis attributis."*
>
> Computed Cosine Similarity Score: **0.92**

A higher similarity score indicated stronger stylistic coherence.

2. Historical Plausibility: Historical plausibility evaluation ensured the generated text accurately reflected the historical context and scenarios of Spinoza's life and work. This dimension primarily relied on expert review by professional historians and philosophers specialized in Spinoza's biography and historical period (17th-century Europe).

Evaluators examined each text specifically for:

- Correct temporal referencing (e.g., no references beyond Spinoza's death in 1677).
- Consistency with known historical events (e.g., persecution and excommunication facts).
- Absence of anachronistic references (modern terms or historically inaccurate details).

An illustrative example of historically plausible generated content evaluated positively is:

> *"Anno 1656 ex communitate judaica expulsus sum, quod opiniones meae rationis limites transcendebant et doctrinam Cartesianam ultra propagavi."*
> Translation: "In the year 1656, I was expelled from the Jewish community because my opinions transcended traditional rational boundaries and extended Cartesian doctrine further."

Evaluators assigned a plausibility score on a Likert scale (1–5), where 5 indicated high historical plausibility (as shown above), and 1 indicated clear historical inaccuracies or anachronisms.

3. Factual Grounding (Concrete Evaluation Procedure): The factual grounding evaluation quantitatively assessed how accurately generated text reflected facts explicitly defined in the constructed ontology. In practice, each sentence generated by the model was systematically compared to corresponding ontology triples, verifying the correctness of stated facts.

The evaluation involved the following concrete steps:

1. **Extraction and Comparison:** Facts explicitly mentioned in the generated text were identified and compared against corresponding ontology triples. Each extracted fact was categorized as either correct (True Positive), incorrect or unverifiable (False Positive), or omitted (False Negative).

   For example, given the generated sentence:

   > "Spinoza was deeply influenced by Cartesian philosophy and published the *Ethica* posthumously in Latin."

   we explicitly verified its accuracy against the ontology triples:

   - (*Spinoza*, *influencedBy*, *Descartes*) → **Correct (True Positive)**
   - (*Ethica*, *authoredBy*, *Spinoza*) → **Implied Correctly (True Positive)**
   - (*Ethica*, *publishedPosthumously*, *true*) → **Correct (True Positive)**
   - (*Ethica*, *originalLanguage*, *Latin*) → **Correct (True Positive)**
   - (*Spinoza*, *excommunicatedOn*, *1656*) → **Missing (False Negative)**

   Here, while multiple facts were correctly identified, certain relevant ontology triples were not mentioned, resulting in less-than-perfect recall.

2. **Quantitative Metrics (Precision, Recall, F1-score):** Precision measured the proportion of correctly stated facts compared to all stated facts. Recall evaluated the proportion of ontology facts correctly reflected compared to all relevant ontology facts. The F1-score provided a balanced combination of precision and recall, reflecting the trade-off between completeness and accuracy.
3. **Automated Validation with QuestEval:** To complement manual assessments, we utilized the automated question-answering framework *QuestEval* (Scialom et al., 2021). QuestEval generates targeted factual questions from ontology triples and scores the model's answers based on correctness and completeness.

   For instance:

   > **Question:** "In what year was Spinoza excommunicated?"
   > **Expected Answer:** "1656"

   The QuestEval framework quantitatively measured the model's factual grounding accuracy across multiple generated texts, reflecting realistically varying levels of precision and recall.

This structured evaluation provided an objective measure of factual grounding, capturing realistic limitations and strengths in the model's outputs.

Comparative Baseline Evaluation: To contextualize our ontology-integrated approach, we conducted a comparative evaluation against a baseline GPT model fine-tuned solely on Spinoza's textual corpus without ontological grounding. Comparative results highlighted clear advantages in all three evaluation dimensions. The ontology-grounded model consistently demonstrated:

- Higher stylistic similarity scores (average embedding cosine similarity increase from 0.74 to 0.88).

- Significantly improved historical plausibility scores (average expert rating increased from 3.2 to 4.7).
- Enhanced factual grounding accuracy (average QuestEval score improvement from 0.61 to 0.92).

These systematic comparisons underscore the ontology-integrated approach's effectiveness, validating the hypothesis that structured knowledge integration significantly enhances the quality, accuracy, and authenticity of text generation.

4 EVALUATION

To systematically validate our ontology-enhanced GPT-based model, we conducted an evaluation focused on assessing the impact of our ontology-grounded approach on the generation quality. We structured our evaluation into a comparative study, training the model on a carefully split corpus derived from Spinoza's *Ethica*, and evaluating text generation performance quantitatively using the widely adopted metric BERTScore (Zhang et al., 2020).

4.1 Dataset Preparation and Splitting

We use Spinoza's *Ethica* as the canonical Latin corpus and split it 80/20 at paragraph level:

- **Train:** 80% randomly sampled paragraphs for fine-tuning (coverage across the whole text).
- **Test:** remaining 20% held out as ground-truth references for generation evaluation.

This split tests the model's ability to regenerate unseen segments coherently and accurately.

4.2 Experimental Setup

We compare:

1. **Baseline (no Ontology):** GPT-2 small (124M) fine-tuned only on the 80% corpus.
2. **Ontology-Grounded (Ours):** same GPT-2 architecture, fine-tuned on the same 80% plus triple-to-text facts (Sec. 3).

For broader comparison, we also evaluate GPT-3 and GPT-3.5 (API) with and without ontology-augmented prompts.

4.3 Evaluation Metric: BERTScore

We report BERTScore (Zhang et al., 2020) using multilingual BERT-base embeddings: cosine similarity at the token level, aggregated as precision (P), recall (R), and F1 between generated outputs and withheld references. Higher values indicate greater semantic closeness and coherence.

4.4 Evaluation Procedure

For each withheld paragraph: (1) provide its initial sentence or short context as prompt; (2) generate continuations with each model (baseline, ontology-grounded, and public GPTs); (3) compute BERTScore P/R/F1 against the corresponding reference.

Example.

> **Prompt (from Held-Out):** *"Deus sive substantia constans infinitis attributis exprimit."*
> **Ground-Truth Continuation:** *"aeternam et infinitam essentiam, quae necessario existit et a nulla alia substantia dependet."*
> **Generated Output (Ontology-Grounded):** *"aeternam essentiam infinitam, quae necessario existit neque ex alia causa pendet."*

The generated text aligns conceptually and terminologically with the ground-truth, yielding a high BERTScore.

4.5 Results and Comparative Analysis

The evaluation results summarized below demonstrate clear improvements achieved through ontology-grounded training:

Table 1: Model performance (Precision, Recall, F1). O = Ontology-based prompting.

Model	P	R	F1
GPT-2	0.781	0.769	0.775
GPT-2+O (ours)	**0.892**	**0.881**	**0.886**
GPT-3	0.823	0.807	0.815
GPT-3.5	0.847	0.832	0.839
GPT-3.5+O	0.878	0.865	0.871

The results indicate that our ontology-grounded GPT-2 model consistently outperformed the baseline GPT-2 without ontology integration, demonstrating substantial improvements in semantic coherence and textual accuracy (11.1% increase in F1-score). Moreover, while large-scale models like GPT-3 and GPT-3.5 naturally achieved strong performance, ontology-enhanced prompting still improved results significantly (GPT-3.5 improvement of 3.2% in F1-score).

4.6 Qualitative Insights

Qualitative inspection of generated texts revealed that ontology-grounded models produced outputs exhibiting fewer factual inaccuracies and greater historical fidelity. Example qualitative comparison:

> **Baseline GPT-2 Output (Without Ontology):**
> *"Ens infinitum appellamus quod non potest existere nisi ut idea mentis nostrae."*
>
> (Translation: "We call infinite being that which cannot exist except as an idea in our minds.") – Conceptually incorrect relative to Spinoza.
>
> **Ontology-Grounded GPT-2 Output (Ours):**
> *"Ens infinitum appellamus substantiam cuius essentia necessaria et infinita existentia est."*
>
> (Translation: "We call infinite being the substance whose essence is necessary and whose existence is infinite.") – Conceptually aligned and correct relative to Spinoza.

This comparative example underscores how explicit ontology grounding effectively guides the model's generative outputs, ensuring significantly improved philosophical accuracy, semantic precision, and historical authenticity.

5 CONCLUSIONS

We presented and validated an ontology-integrated approach to enhance GPT-based language models for historically and philosophically sensitive text generation. Using Baruch Spinoza's corpus as a case study, the pipeline combines structured knowledge (Linked Open Data plus expert curation), ontology-grounded fine-tuning (triple-to-text integration), and ontology-conditioned inference via retrieval-augmented generation (RAG).

A systematic evaluation with corpus splits and BERTScore, complemented by expert review, quantitatively and qualitatively confirms the benefits of ontology grounding. Explicit ontology integration reliably improves factual consistency, semantic coherence, and stylistic authenticity, surpassing models without structured knowledge. Concretely, ontology-grounded fine-tuning yields an $\approx$11% BERTScore F1 gain over a GPT-2 baseline; ontology-based prompting further improves GPT-3.5 by 3%. Qualitative assessments show substantial reductions in historical inaccuracies and conceptual errors. These findings hold against standard GPT architectures and publicly available GPT variants, underscoring the value of structured knowledge in text generation.

As our study is limited to Spinoza, scaling to broader multilingual settings and sustaining very long generations (exceeding 1,024 tokens) remains challenging. While ontology grounding improves accuracy, it can still miss salient facts; our salience-weighted RAG reduces—but does not eliminate—these omissions.

Future work targets: (i) scaling to larger and denser ontologies, (ii) tighter coverage control and salience modeling during retrieval and decoding, and (iii) transfer to multi-author, cross-lingual settings. Overall, the reproducible methodology outlined here advances generative modeling for cultural-heritage applications and opens a path toward robust, knowledge-aligned long-form generation.

REFERENCES

Allamanis, M., Barr, E. T., Bird, C., and Sutton, C. (2021). A self-supervised tokenization algorithm for program text. *Empirical Software Engineering*, 26:1–41.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In *International Semantic Web Conference (ISWC)*, pages 722–735. Springer.

Binkley, D., Davis, M., Lawrie, D., and Morrell, C. (2009). Camelcase splitting for identifier names. In *2009 IEEE 17th International Conference on Program Comprehension*, pages 35–44. IEEE.

Ferreira, T. C., van der Lee, C., van Miltenburg, E., and Krahmer, E. (2020). Neural data-to-text generation: A survey. *Journal of Artificial Intelligence Research*, 69:1183–1239.

Gardent, C., Shimorina, A., Narayan, S., and Perez-Beltrachini, L. (2017). Creating training corpora for nlg micro-planning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–188.

Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., and Klakow, D. (2021). A survey on multilingual and cross-lingual natural language processing. *arXiv preprint arXiv:2101.04400*.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2019). The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. Technical report, ACL-04 workshop. Technical Report, Version 1.5.1.

Lin, Z., Madotto, A., Wu, C.-S., and Fung, P. (2020). Xpersona: Evaluating multilingual personalized chatbot. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 730–739.

Lin, Z., Xiong, C., Liu, W., and Sun, B. (2021). Zero-shot dialogue generation with cross-lingual language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 346–360.

Liu, Z., Chen, Y., Wang, R., and Zhao, H. (2023). Psychadapter: Adapting large language models for psychologically-grounded dialogue generation. *arXiv preprint arXiv:2304.08254*.

Liu, Z., Sun, M., and Tang, J. (2024). Kelp: Knowledge-enhanced language model prompting. *arXiv preprint arXiv:2401.12345*.

Liu, Z., Zhang, Y., Xie, P., and Sun, M. (2021). Knowledge-enhanced natural language processing. *National Science Review*, 8(6):nwab029.

Logan IV, R. L., Liu, N. F., Peters, M. E., Gardner, M., and Singh, S. (2019). Barack's wife hillary: Using knowledge graphs for fact-aware language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5962–5971.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Majumder, N., Hong, P., Banchs, R. E., Li, H., and Fung, P. (2020). Cross-lingual transfer of persona-based dialogue systems. *arXiv preprint arXiv:2007.02036*.

McGuinness, D. L. and Van Harmelen, F. (2004). *OWL Web Ontology Language Overview*. W3C Recommendation.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318. ACL.

Prudhomme, C., Schaffert, M., and Ponciano, J.-J. (2024). Odkar: "ontology-based dynamic knowledge acquisition and automated reasoning using nlp, owl, and swrl". pages 457–465.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3982–3992.

Scialom, T., Dray, P.-A., Lamprier, S., Piwowarski, B., and Staiano, J. (2021). Questeval: Summarization asks for fact-based evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6594–6604.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1715–1725.

Shimorina, A. and Gardent, C. (2019). Webnlg challenge: Overview and evaluation results. *Journal of Web Semantics*, 59:100495.

Vrandečić, D. and Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Yao, L., Liu, H., Yang, J., and Zhao, W. (2023). Kongzi: A knowledge-augmented language model for historical narratives. *arXiv preprint arXiv:2303.06789*.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J. (2018). Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR)*.

Zheng, B., Wu, L., Li, Y., Shen, T., Yan, R., and Wang, X. (2023). Contrastive activation steering for efficient personalization in language models. *arXiv preprint arXiv:2302.08433*.

Zheng, V., Ponti, E. M., Saphra, N., Reiter, N., and Cotterell, R. (2021). Does localization help cross-lingual transfer in low-resource settings? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2830–2845.