# From Graph to Text and Back: Semantic Fidelity in Automated Industrial Knowledge Graphs

**Kamyar Zeinalipour**[1], **Silvia Severini**[2], **Alessia Borghini**[2,*], **Sara Cardarelli**[2,*], **Marco Maggini**[1], **Marco Gori**[1]
[1]DIISM - University of Siena, Italy
[2]Leonardo - Rome, Italy
kamyar.zeinalipour2@unisi.it

## Abstract

Knowledge Graphs (KGs) are the backbone of reliable industrial data strategies, yet verbalizing them with Large Language Models (LLMs) often leads to unacceptable risks for high-stakes applications, such as hallucinations or omitted relations. To enforce strict semantic fidelity in KG-to-text generation, we introduce a self-supervised *round-trip* pipeline. The system verbalizes KG triples into text and immediately attempts to reconstruct the original graph from that text; only verbalizations that enable perfect graph recovery are retained. This creates a closed feedback loop that guarantees the generated text is semantically equivalent to the source data. Experiments confirm that our automated round-trip consistency score correlates strongly with expert judgment, effectively acting as a scalable proxy for human review. Furthermore, we show that standard LLMs can bootstrap their own KG-extraction and generation capabilities by fine-tuning on this trusted synthetic data. Our approach yields significant improvements in triple-extraction accuracy and verbalization faithfulness without relying on costly manual annotation or massive teacher models, offering a practical path to deploying trustworthy, KG-grounded AI systems.

## 1 Introduction

In the enterprise landscape, Knowledge Graphs (KGs) serve as the backbone of reliable data strategies, providing a structured source of truth for downstream applications (Shao and Kumar, 2022; Jiang et al., 2023; Osuji et al., 2024). However, bridging the gap between this rigid structured data and user-friendly natural language remains a critical bottleneck for industrial adoption (Li et al., 2022). While Large Language Models (LLMs) offer fluency, they introduce a "black box" risk: state-of-the-art models frequently hallucinate entities or omit relations, posing unacceptable liability risks in precision-critical domains such as healthcare, finance, and industrial process monitoring (Rebuffel et al., 2022b; Ji et al., 2023; Zeinalipour et al., 2025). For industry, the challenge is not merely generating text, but guaranteeing its **semantic fidelity** without incurring the prohibitive cost of manual verification.

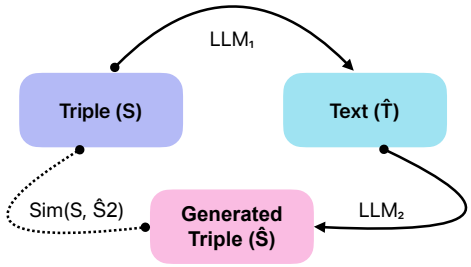


Figure 1: Overview of our proposed two-stage round-trip pipeline. Structured inputs ($S$) are verbalized into candidate texts ($\hat{T}$) (Stage 1), then reconverted to structured form ($\hat{S}$) (Stage 2). A similarity measure $\mathrm{sim}(S, \hat{S}, \alpha)$ evaluates the transformation quality.

**Problem statement.** To address this reliability gap, we formulate the following question:

*How can we generate synthetic KG-to-text data that is guaranteed to be semantically equivalent to the source, enabling models to learn from their own output without human supervision?*

Solving this yields two distinct industrial advantages: (1) the creation of high-fidelity synthetic corpora without costly manual intervention, and (2) the establishment of a scalable, automated proxy for expensive Subject Matter Expert (SME) quality judgments.

**Approach overview.** We propose a two-stage round-trip pipeline (Figure 1). Stage 1 verbalizes a structured input $S$ into a candidate text $\hat{T}$; Stage 2 maps $\hat{T}$ back to a structured form $\hat{S}$. We compare the pair using a *round-trip similarity* $\mathrm{sim}(S, \hat{S}, \alpha)$

*Work was done while at Leonardo.



*Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 2095–2111

combining embedding-based semantics and lexical overlap. A **dynamic-sampling controller** iteratively adjusts the decoding parameters to retain only candidates exceeding a target similarity. We illustrate this process in Figure 2 using a case study on the Komodo Dragon. We investigate:

**RQ1** Does $\mathrm{sim}(S, \hat{S}, \alpha)$ reliably track human ratings, serving as a scalable automated evaluator to replace manual review?

**RQ2** Can a model bootstrap its extraction accuracy by fine-tuning on its own high-fidelity synthetic text, thereby reducing the need for massive teacher models or labeled data?

**Key findings.** Across our 250-graph corpus, we find that: (i) the automatic score $\mathrm{sim}(S, \hat{S}, \alpha)$ aligns closely with human judgments, validating its use as a quality gate; (ii) self-training on this verified corpus boosts extraction quality on *both* automatic metrics and blinded human studies, demonstrating a viable path for self-improving systems.

**Contributions.** In summary, our contributions are: (1) an architecture-agnostic round-trip pipeline for producing *high-fidelity* synthetic text from structured enterprise assets; (2) a similarity-driven dynamic-sampling mechanism that automatically calibrates decoding to ensure maximum semantic retention; (3) evidence that models can *self-bootstrap* extraction performance via this pipeline, outperforming distillation baselines without manual labeling; and (4) the public release of code, checkpoints, and datasets.[1]

## 2 Related Work

**Risk Mitigation in Generative Systems** Neural data-to-text systems frequently suffer from hallucinations and critical omissions (Wiseman et al., 2017; Mei et al., 2016; Dušek et al., 2019; Thomson and Reiter, 2020; Rebuffel et al., 2022a). Evaluation typically relies on n-gram overlaps (Papineni et al., 2002; Banerjee and Lavie, 2005; Lin, 2004) or coverage-aware metrics like PARENT (Dhingra, 2019). To ensure fidelity, prior work employs critic models (Harkous et al., 2020; Lango and Dušek, 2023), cycle-consistency losses (Guo, 2020; Gong, 2020), or specialized decoding strategies (Holtzman, 2020). Our work operationalizes these insights via an automated back-parsing pipeline to guarantee high-fidelity outputs.

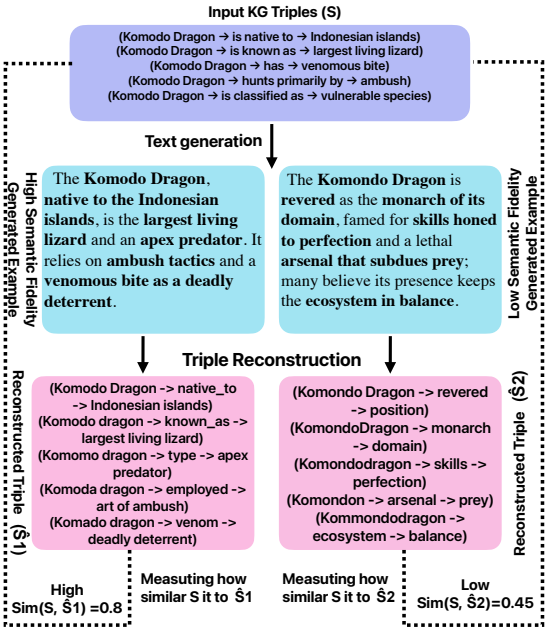


Figure 2: The input KG triples about the "Komodo Dragon" are first verbalized by the generator (Stage 1). The reconstructor then maps the text back to triples (Stage 2). High-fidelity text (left) reproduces every fact, whereas a low-fidelity variant (right) introduces omissions and errors, leading to low round-trip similarity.[2]

**Scalable Data Synthesis & Bootstrapping** Synthetic data reduces annotation costs. To ensure high-quality data, round-trip methods have been used, such as back-translation (Sennrich et al., 2016; He, 2016) or unsupervised simplification (Surya, 2019; Schumann, 2020). In the LLM era, "born-again" networks and self-distillation (Furlanello, 2018; Hinton, 2015) leverage self-instruction (Wang, 2022b; Long, 2024) and modern language capabilities (Radford et al., 2019; Liu et al., 2024; Long et al., 2024; Bauer et al., 2024) to bootstrap performance across diverse domains, from structured data extraction to protein sequence design (Zeinalipour et al., 2024b). Unlike teacher-dependent methods, we employ a self-contained loop filtering samples via embedding similarity (Wang, 2022a) to ensure semantic integrity.

**Knowledge Graph-to-Text Systems** KG-to-text has evolved from rule-based pipelines (Reiter, 1997; Dale and Reiter, 1998) to neural models, which often trade fidelity for fluency (Castro Ferreira, 2018; Mei et al., 2016; Wiseman et al., 2017). Solutions range from specialized graph en-

[1]Code and datasets available at https://github.com/KamyarZeinalipour/round-trip-kg under MIT license.

[2]We compressed the generated texts of Figure 2 to enhance visualization; the original text appears in the Appendix A.

coders (Marcheggiani and Perez-Beltrachini, 2018; Koncel-Kedziorski et al., 2019) and explicit planners (Puduppully et al., 2019; Moryossef, 2019) to pre-trained integrations like JointGT (Ke, 2021) and MGSA (Wang et al., 2024). Recent work emphasizes few-shot prompting (Li, 2021; Zhang et al., 2024; Zhao, 2023) and large synthetic datasets (Kim et al., 2024). Parallel efforts in educational NLP have applied similar structured-knowledge-to-text paradigms to generate pedagogical content from knowledge bases (Zeinalipour et al., 2024a,c). We complement these with a model-agnostic verification layer adaptable to standard LLMs.

## 3 Methodology

We propose a domain-agnostic **structure-to-text-and-back loop** for generating synthetic text that faithfully retains input semantics. Demonstrated here with KG triples, the method generalizes to any structured representation as it relies on semantic consistency rather than task-specific constraints. The pipeline (Figure 3) comprises four steps: (1) Structure-to-Text Generation, (2) Text-to-Structure Reconstruction, (3) Iterative Refinement via Dynamic Sampling, and (4) Automated Selection. Algorithm 1 details this adaptive process.

### 3.1 Workflow Overview

Given structured input $S$ (e.g., KG triples), the pipeline performs $N$ iterative cycles to generate text $\hat{T}$ and a structured reconstruction $\hat{S}$. Generation parameters are dynamically adjusted via the *round-trip similarity*[3]. The embedding function $\phi(\cdot)$ is implemented using `multilingual-e5-large-instruct` (Wang, 2022a), a state-of-the-art multilingual sentence encoder chosen for its ability to capture deep structural semantics regardless of surface phrasing—a critical requirement when comparing structured KG tuples against fluid natural language:

$$\mathrm{sim}(S, \hat{S}, \alpha) = \alpha \cdot \cos\big(\phi(S), \phi(\hat{S})\big) + (1-\alpha) \cdot \mathrm{sim}_{\text{lex}}(S, \hat{S})$$

where $\phi(\cdot)$ represents embeddings, $\mathrm{sim}_{\text{lex}}$ is a lexical metric, and $\alpha \in [0,1]$ balances semantic and lexical alignment.

The workflow proceeds as follows:

1. **Structure-to-Text Generation:** $\text{LLM}_1$ verbalizes structured data ($S$) into candidates ($\hat{T}$). To ensure diversity, subsequent iterations ($i \geq 3$) explicitly prompt the model to produce text distinct from previous versions, while the initial pass ($i = 1$) generates directly.

2. **Text-to-Structure Reconstruction:** A complementary $\text{LLM}_2$ converts $\hat{T}$ back into structured format $\hat{S}$ to measure the semantic preservation of the transformation.

3. **Iterative Refinement:** We evaluate fidelity via round-trip similarity and dynamically adjust decoding temperature ($t$) and top-p ($p$) to balance consistency and diversity:

$$(t, p) \leftarrow (t, p) + \begin{cases} (-\Delta t, -\Delta p), & \mathrm{sim}(S, \hat{S}, \alpha) < \tau_{\text{low}}; \\ (0, 0), & \tau_{\text{low}} \leq \mathrm{sim}(S, \hat{S}, \alpha) < \tau_{\text{high}}; \\ (+\Delta t, +\Delta p), & \mathrm{sim}(S, \hat{S}, \alpha) \geq \tau_{\text{high}}. \end{cases}$$

   Thresholds and increments are tuned on the development set. This mechanism reduces randomness to recover fidelity when scores are low, and increases it to promote diversity when scores are high.

4. **Automated Selection:** After $N$ cycles, outputs surpassing a strict threshold $\tau_{select}$[4] are automatically retained, yielding a high-fidelity synthetic dataset without human intervention.

## 4 Experiments

We empirically assess the proposed *structure-to-text-and-back loop* via two research questions (reproducibility details in Appendix B):

**RQ1:** *Does round–trip similarity* $\mathrm{sim}(S, \hat{S}, \alpha)$ *track human judgments of text quality?*
We generated synthetic text across diverse fidelity levels (low, medium, high) using our loop and correlated the automated round-trip scores with human ratings of the outputs.

**RQ2:** *Can a model bootstrap its own performance on KG-extraction by fine-tuning on self-created high-fidelity data?*

[3] Hyperparameters detailed in Appendix B.4.

[4] Determined via grid-search on the dev set.

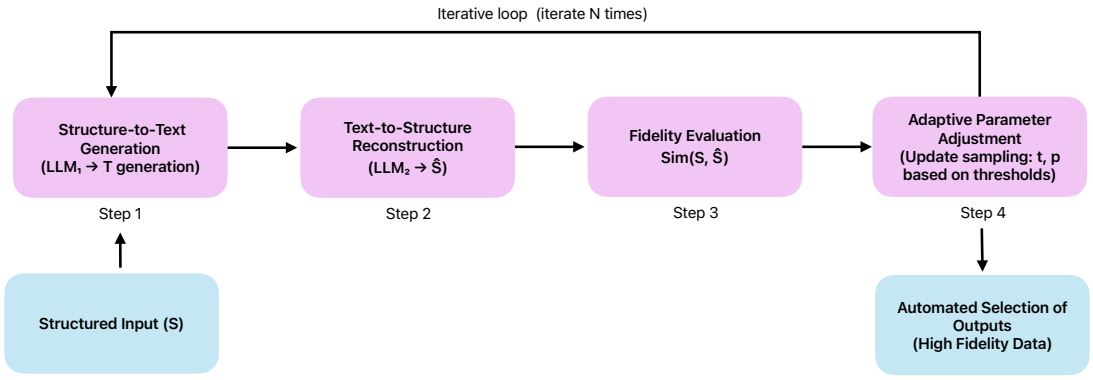


Figure 3: Illustration of the iterative "structure-to-text-and-back" loop. **Step 1 (Generation):** Structured inputs are converted into candidate texts by $\text{LLM}_1$. **Step 2 (Reconstruction):** $\text{LLM}_2$ maps these texts back into structured form. **Step 3 (Refinement):** Semantic and lexical similarity scores between the original and reconstructed data drive the adaptive adjustment of decoding parameters (temperature $t$, top-$p$). **Step 4 (Selection):** Only outputs exceeding a high-fidelity threshold are retained. The cycle repeats $N$ times to optimize synthetic quality.

We selected the highest-fidelity synthetic examples from our pipeline and fine-tuned three LLMs to reconstruct KGs from text. Performance was measured via both automated metrics and human evaluation.

**Models.** We employed three open-weights models to bracket the small/medium parameter regime: `LLaMA-2-7B`, `LLaMA-3.2-3B`, and `LLaMA-3.2-1B` (Touvron et al., 2023a,b).

**Seed KGs.** Our dataset comprises 250 KGs derived from diverse English Wikipedia topics. Annotators manually created five triples per subject, which were subsequently verified by an independent group. The data is partitioned into:

**Training Set (200 KGs):** Used for synthetic generation (RQ1) and self-training (RQ2) (see Appendix D.1).

**Evaluation Set (50 KGs):** Paired with human-written paragraphs to test extraction performance (RQ2) (see Appendix D.2).

### 4.1 Evaluation Protocol

Three postgraduate linguistics students (IELTS 7) evaluated the dataset quality, following the guidelines in Appendix C.

**Human evaluation for RQ1.** Annotators scored 250 generated paragraphs (100 overlapping for agreement) on a five-point A–E scale. The primary metric was **Content & Relation Accuracy (CRA)**: does the text faithfully reproduce triples without hallucination? We also logged **Structure, Grammar & Fluency (SGF)** and **Originality, Engagement & Creativity (OEC)** to analyze how linguistic quality correlates with fidelity.

| Dimension | % Agreement | Fleiss's $\kappa$ |
|---|---|---|
| CRA | 89.8 | 0.63 |
| SGF | 87.6 | 0.61 |
| OEC | 88.2 | 0.68 |

Table 1: Inter-annotator agreement on the 100 shared items.

**Human evaluation for RQ2.** Annotators performed a blinded side-by-side comparison of triples extracted by the base vs. fine-tuned models. They assigned a consolidated quality rating (A–E) covering template adherence, meaning preservation, and structure, and indicated a model preference. Each annotator evaluated 200 sets (100 overlapping).

**Automatic metrics.** For RQ1, we report Pearson's $r$ between $\text{sim}(S, \hat{S}, \alpha)$ and human scores. For RQ2, we evaluate extraction quality against the 50 gold-standard KGs using BERTSCORE (with `XLM-RoBERTa-base`) (Zhang et al., 2019), ROUGE-L (Lin, 2004), and BLEU (Papineni et al., 2002).

### 4.2 Results and Analysis

**RQ1 – Validity of the Automatic Metric.** To validate the metric independently, we sampled generated outputs *before* applying any selection filter ($\tau_{\text{select}}$), explicitly covering the full fidelity spectrum: early-iteration failures and severe hal-

**Algorithm 1** Structure-to-Text-and-Back Loop

**Require:** Structured data ($S$), models ($\text{LLM}_1$, $\text{LLM}_2$), iterations ($N$), temperature ($t_0$), top-p ($p_0$), increments ($\Delta t, \Delta p$), fidelity thresholds ($\tau_{\text{low}}, \tau_{\text{high}}, \tau_{\text{select}}$), balance parameter $\alpha$

selected $\leftarrow \{\}$
$t, p \leftarrow t_0, p_0$
previousText $\leftarrow \perp$
**for** $i \leftarrow 1$ **to** $N$ **do**
  ▷ **Step 1: Structure-to-Text Generation**
  **if** $i = 1$ **then**
    $\hat{T} \leftarrow \text{LLM}_1$.generate_text($S$; $t$, $p$)
  **else**
    $\hat{T} \leftarrow \text{LLM}_1$.generate_text($S \parallel$ previousText; $t$, $p$)
  **end if**
  ▷ **Step 2: Text-to-Structure Reconstruction**
  $\hat{S} \leftarrow \text{LLM}_2$.extract_structure($\hat{T}$)
  ▷ **Step 3: Fidelity Evaluation**
  fidelity $\leftarrow \text{sim}(S, \hat{S}, \alpha)$
  ▷ **Step 4: Adaptive Parameter Adjustment & Selection**
  **if** fidelity $< \tau_{\text{low}}$ **then**
    $t \leftarrow \max(0, t - \Delta t)$
    $p \leftarrow \max(0, p - \Delta p)$
  **else if** fidelity $\geq \tau_{\text{high}}$ **then**
    $t \leftarrow \min(1, t + \Delta t)$
    $p \leftarrow \min(1, p + \Delta p)$
  **end if**
  **if** fidelity $\geq \tau_{\text{select}}$ **then**
    selected $\leftarrow$ selected $\cup \{(S, \hat{T})\}$
  **end if**
  previousText $\leftarrow \hat{T}$
**end for**
**return** selected

lucinations (low scores), partially correct outputs (medium), and high-quality generations (high). Human evaluators then *blindly* rated this unfiltered set, ensuring no circular dependency between metric selection and evaluation. We converted A–E ratings to a 5–1 scale. Figure 4 shows strong positive correlations (Pearson's $r > 0.5$) between our automatic score $\text{sim}(S, \hat{S}, \alpha)$ and human ratings for CRA and SGF across this full unfiltered range, validating the metric as an effective, independent, low-cost quality filter. Correlation with OEC was positive but lower, indicating that fidelity scores only partially capture creativity. Inter-annotator agreement (Table 1) was substantial (87.6%–89.8%, Fleiss's $\kappa$ 0.61–0.68), confirming the reliability of human assessments.

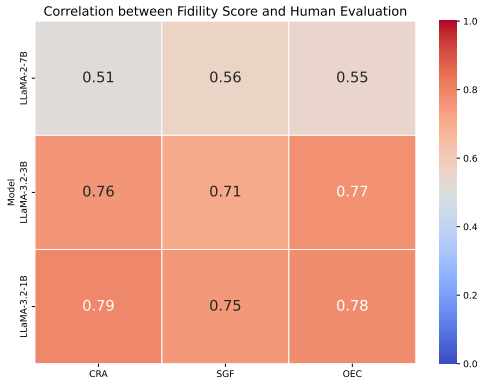


Figure 4: Heatmap visualization of the correlation between the automated round-trip similarity and human evaluation dimensions (CRA, SGF, OEC). Darker colors indicate stronger correlations.

| Model | Off-the-Shelf | Fine-Tuned | Δ |
|---|---|---|---|
| LLaMA2-7B | 2.70 | 3.32 | +0.62 |
| LLaMA3.2-3B | 2.30 | 4.02 | +1.72 |
| LLaMA3.2-1B | 1.94 | 2.70 | +0.76 |

Table 2: Average human ratings comparing off-the-shelf and fine-tuned model variants. Higher scores indicate better performance; Δ is the improvement of the fine-tuned model over the off-the-shelf counterpart.

**RQ2 – Self-Synthetic Fine-Tuning.** We fine-tuned `LLaMA-3.2-1B`, `3.2-3B`, and `2-7B` to generate KGs (five triples) from paragraphs using only self-generated data. Using our loop (100 cycles), we selected the top-10 high-fidelity texts for each of the 200 training KGs, yielding 2,000 synthetic training samples. Testing was performed on the 50 held-out KG-text pairs (Sec. 4; examples in Appendix E). Human annotators rated base vs. fine-tuned models on a 5-point scale (converted A–E → 5–1). Table 2 confirms that fine-tuned models consistently outperform base versions. Table 1 shows substantial agreement on the 100 shared items across CRA (89.8%, $\kappa = 0.63$), OEC (88.2%, $\kappa = 0.68$), and SGF (87.6%, $\kappa = 0.61$), confirming assessment reliability. In blinded side-by-side comparisons, annotators strongly preferred the fine-tuned variants (Figure 5). Automated metrics (BERT F1, ROUGE-L F1, BLEU-4) further corroborate these improvements (Table 3). We note that absolute automatic metric values (e.g., BLEU-4 peaking at 0.151) may appear modest; however, this reflects well-documented limitations of n-gram

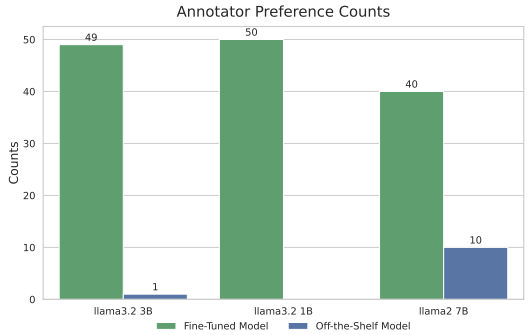


Figure 5: Annotator preferences indicating strong favorability towards fine-tuned models compared to off-the-shelf versions.

and string-matching metrics when applied to structured extraction tasks rather than model failure. Multiple valid triple formulations exist for the same semantic content (e.g., "Parthenon → located_in → Athens" vs. "Parthenon → city → Athens"), and minor lexical differences are heavily penalized by these metrics despite semantic equivalence. This is precisely why blinded human evaluation was adopted as our primary standard, where the fine-tuned 3B model achieves a +1.72 point improvement and is overwhelmingly preferred by annotators (49-to-1, Figure 5).

The rating distribution in Figure 6 highlights that fine-tuning shifts outputs toward the higher-quality end of the scale. Finally, Figure 7 demonstrates that performance gains are monotonic with respect to synthetic data volume, with significant jumps appearing even at 25% data usage.

**Efficiency Ablation.** To evaluate the impact of data volume, we retrained models using 25%, 50%, 75%, and 100% of the synthetic dataset, averaging results over three random seeds per fraction. As shown in Figure 7, the most significant performance jumps occur at just 25%, with diminishing returns thereafter. This demonstrates that *quality* drives improvement more than sheer quantity, suggesting that small, high-fidelity datasets can efficiently boost structured knowledge extraction while minimizing resource costs.

## 5 Conclusion

Deploying trustworthy, high-stakes industrial AI requires accurately verbalizing structured enterprise assets. To affordably overcome the "hallucination bottleneck," our **high-fidelity round-trip pipeline** enforces strict semantic consistency via a self-supervised feedback loop, which experiments validate for industrial deployment:

**RQ1 (Automated Auditing):** We demonstrated that our automated **round-trip similarity score** $\text{sim}(S, \hat{S}, \alpha)$ correlates strongly with expert human judgment ($r > 0.5$). This confirms the metric can serve as a scalable, automated proxy for costly subject-matter expert review. **RQ2 (Cost-Effective Bootstrapping):** We proved that open-weights LLMs can significantly improve their own extraction capabilities using *only* self-generated data. Notably, `LLaMA3.2-3B` achieved a human-rated quality jump of **+1.72 points** (on a 5-point scale), demonstrating that expensive proprietary teacher models are not required for high performance. Crucially, ablation studies revealed that these gains are driven by data *fidelity* rather than volume, with substantial improvements appearing after using only 25% of the synthetic corpus. This offers a highly efficient path for deploying reliable, KG-grounded AI systems while minimizing compute and labeling overhead. Future work will extend this architecture to complex enterprise formats, such as hierarchical databases and logs.

## Limitations

While our proposed method demonstrates significant advantages, several limitations merit discussion.

**Experimental scope.** Our experiments use KG triples with five triples per subject. Although the pipeline is representation-agnostic—since similarity evaluation relies on text-to-text semantic comparisons rather than graph-specific heuristics—its practical efficacy on more complex industrial structured data (e.g., deeply nested hierarchical databases, enterprise logs, or irregular schemas with domain-specific long-tail entities) remains to be rigorously evaluated. We note that chunking KGs into 5-triple subgraphs mirrors the localized subgraph retrieval step used in real-world industrial GraphRAG pipelines, which do not feed massive, unchunked graphs to LLMs. Nevertheless, scaling experiments to larger and more complex enterprise KGs is an important direction for future work.

**Corpus size.** Our evaluation corpus of 250 Wikipedia-derived KGs is deliberately small to enable meticulous, triple-by-triple expert human evaluation—a process that is prohibitively expensive to scale, and precisely the bottleneck our pipeline addresses. While our ablation study

| Model | BERTScore F1 | | ROUGE-L F1 | | BLEU-4 | |
|---|---|---|---|---|---|---|
| | Base | Fine-Tuned | Base | Fine-Tuned | Base | Fine-Tuned |
| LLaMA2-7B | 0.438 | 0.505 | 0.382 | 0.419 | 0.131 | **0.151** |
| LLaMA3.2-3B | 0.372 | **0.514** | 0.409 | **0.442** | 0.089 | 0.147 |
| LLaMA3.2-1B | 0.329 | 0.399 | 0.347 | 0.300 | 0.072 | 0.086 |

Table 3: Comparison of automated evaluation metrics for off-the-shelf (Base) versus fine-tuned variants. Higher scores indicate better performance.

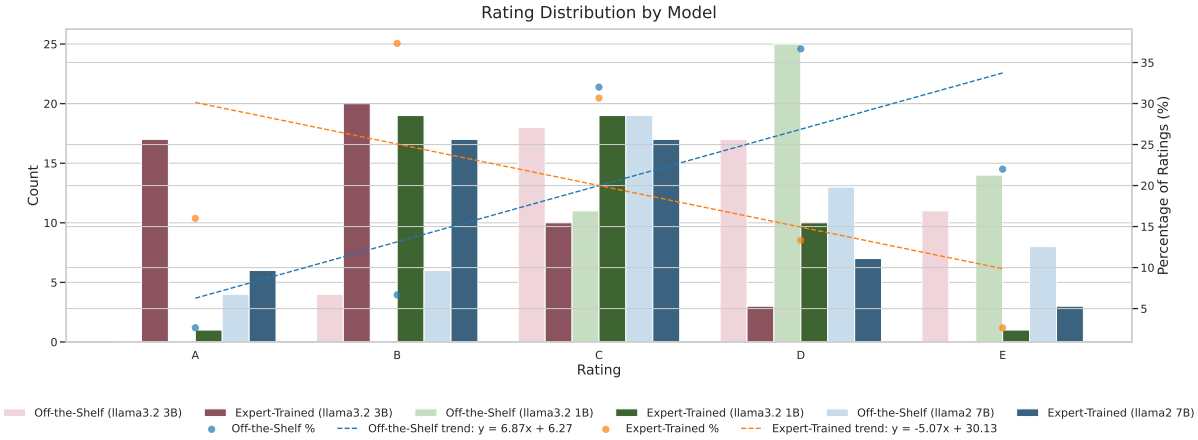


Figure 6: Distribution of annotator ratings (A–E). Darker bars (fine-tuned) dominate the high-quality end (left), while lighter bars (off-the-shelf) concentrate at the low-quality end (right). Dashed trend lines (fine-tuned: $y = -5.07x + 30.13$, off-the-shelf: $y = 6.87x + 6.27$) quantify this divergence.

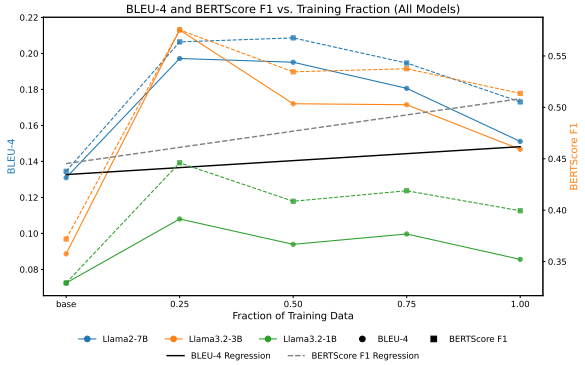


Figure 7: BLEU-4 and BERTScore F1 vs. Fraction of Training Data. Performance rises steadily from 25% to 100%, showing monotonic gains in lexical and semantic quality as synthetic volume increases.

(Figure 7) demonstrates that peak performance is achievable with just 25% of this data (prioritizing quality over quantity), the generalization to highly domain-specific or low-resource KGs with entities poorly represented in LLM pre-training data remains an open question.

**Computational overhead.** Running the generation loop for $N = 100$ iterations per sample incurs a one-time offline compute cost during synthetic dataset construction. However, this cost is strictly offline; once the fine-tuning dataset is created, model training is fast (e.g., ∼15 minutes for the 3B model on our hardware) and the deployed model performs standard single-pass inference with zero additional overhead.

**Self-training risks.** While the structural back-translation step acts as a strong regularizer against factual hallucinations, stylistic biases present in the base model may still propagate through iterative self-training. Future studies should investigate potential saturation points and explore hybrid strategies, such as injecting lightweight expert spot-checks on a small fraction of accepted data before fine-tuning.

**Model scale.** Due to computational constraints, we restricted experiments to models up to 7B parameters. While this allowed clearer observation of quality differences and closer alignment with human judgment, larger models may yield further improvements and should be explored in future work.

## Ethics Statement

LLMs are known to exhibit biases present in their training data and to generate hallucinated content.

We use LLMs for the purpose for which they were designed—generating text and triples—and further train them on self-generated content. Therefore, there is a risk of propagating the intrinsic limitations of LLMs.

While our strict structural back-translation step mitigates factual hallucinations, it does not fully address stylistic or representational biases that may be present in the base models. In industrial settings, we recommend treating our pipeline as a "human-in-the-loop multiplier": injecting lightweight Subject Matter Expert (SME) spot-checks at the Automated Selection phase (Step 4) on a small fraction of accepted data before fine-tuning. This strategy prevents bias drift while still reducing manual annotation costs by over 90%.

The data used in this study was solely utilized as a textual corpus derived from publicly available Wikipedia content, and the content should not be interpreted as an endorsement by our team.

## Acknowledgments

We are grateful to Marina Geymonat and Alessandro Nicolosi for their guidance and for fostering the collaboration that made this work possible. This work is part of the Villanova project, partially supported by IPICEI-CIS, Prog. n. SA. 102519 – CUP B29J24000850005..

## References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation. In *ACL Workshop*.

André Bauer, Simon Trapp, Michael Stenger, Robert Leppich, Samuel Kounev, Mark Leznik, Kyle Chard, and Ian Foster. 2024. Comprehensive exploration of synthetic data generation: A survey. *arXiv preprint arXiv:2401.02524*.

Thiago et al. Castro Ferreira. 2018. Lessons learned from combining principal methods for natural language generation. In *INLG*.

Robert Dale and Ehud Reiter. 1998. Building natural language generation systems. *Journal of Natural Language Engineering*.

Bhuwan et al. Dhingra. 2019. Handling divergent reference texts when evaluating table-to-text generation. In *ACL*.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2019. Semantic noise affects the quality of data-to-text training data. In *INLG*.

Tommaso et al. Furlanello. 2018. Born-again neural networks. In *ICML*.

Heng et al. Gong. 2020. Tablegpt: Few-shot table-to-text generation with table structure reconstruction and content matching. In *COLING*.

Qipeng et al. Guo. 2020. Cyclegt: Unsupervised graph-to-text and text-to-graph generation via cycle training. In *WebNLG Workshop*.

Hamza Harkous, Isabel Groves, and Amir Saffari. 2020. Have your text and use it too! end-to-end neural data-to-text generation with semantic fidelity. In *COLING*.

Di et al. He. 2016. Dual learning for machine translation. In *NeurIPS*.

Geoffrey et al. Hinton. 2015. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Ari et al. Holtzman. 2020. The curious case of neural text degeneration. In *ICLR*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Structgpt: A general framework for large language model to reason over structured data. *arXiv preprint arXiv:2305.09645*.

Pei et al. Ke. 2021. Jointgt: Graph-text joint representation learning for text generation from knowledge graphs. In *ACL Findings*.

Daehee Kim, Deokhyung Kang, Sangwon Ryu, and Gary Geunbae Lee. 2024. Ontology-free general-domain knowledge graph-to-text generation dataset synthesis using large language model. *arXiv preprint arXiv:2409.07088*.

Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. 2019. Text generation from knowledge graphs with graph transformers. In *NAACL*.

Mateusz Lango and Ondrej Dušek. 2023. Critic-driven decoding for mitigating hallucinations in data-to-text generation. In *EMNLP*.

Junyi et al. Li. 2021. Few-shot knowledge graph-to-text generation with pretrained language models. In *ACL Findings*.

Wei Li, Wenhao Wu, Moye Chen, Jiachen Liu, Xinyan Xiao, and Hua Wu. 2022. Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods. *arXiv preprint arXiv:2203.05227*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL Workshop*.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. 2024. Best practices and lessons learned on synthetic data. In *First Conference on Language Modeling*.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On llms-driven synthetic data generation, curation, and evaluation: A survey. *arXiv preprint arXiv:2406.15126*.

Lin et al. Long. 2024. On llms-driven synthetic data generation, curation, and evaluation: A survey. In *ACL Findings*.

Diego Marcheggiani and Laura Perez-Beltrachini. 2018. Deep graph convolutional encoders for structured data-to-text generation. In *INLG*.

Hongyuan Mei, Mohit Bansal, and Matthew Walter. 2016. Data-to-text generation with macro planning. In *ACL*.

Amit et al. Moryossef. 2019. Step-by-step: Separating planning from realization in neural data-to-text generation. In *NAACL*.

Chinonso Cynthia Osuji, Thiago Castro Ferreira, and Brian Davis. 2024. A systematic review of data-to-text nlg. *arXiv preprint arXiv:2402.08496*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with content selection and planning. In *AAAI*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners.

Caroline Rebuffel, Carolina Scarton, and Ehud Reiter. 2022a. Controlling numerical hallucinations in data-to-text generation. In *EMNLP*.

Clément Rebuffel, Marco Roberti, Laure Soulier, Geoffrey Scoutheeten, Rossella Cancelliere, and Patrick Gallinari. 2022b. Controlling hallucinations at word level in data-to-text generation. *Data Mining and Knowledge Discovery*, pages 1–37.

Ehud et al. Reiter. 1997. Generating summaries of patient records. In *AMIA*.

Raphael et al. Schumann. 2020. Discrete-event sequence to text with cyclical generative adversarial networks. In *COLING*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *ACL*.

Yutong Shao and Arun Kumar. 2022. Structured data representation in natural language interfaces. *A Quarterly bulletin of the Computer Society of the IEEE Technical Committee on Data Engineering*, 45(3).

Shashi Narayan et al. Surya. 2019. Unsupervised sentence simplification using monolingual corpora in multiple languages. In *EMNLP*.

Craig Thomson and Ehud Reiter. 2020. Hallucination and omission errors in neural nlg. *Computational Linguistics*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Liang et al. Wang. 2022a. Text embeddings by weakly-supervised contrastive pre-training. arXiv preprint arXiv:2212.03533.

Shanshan Wang, Chun Zhang, and Ning Zhang. 2024. Mgsa: Multi-granularity graph structure attention for knowledge graph-to-text generation. In *2024 12th International Conference on Information Systems and Computing Technology (ISCTech)*, pages 1–5. IEEE.

Yizhong et al. Wang. 2022b. Aligning language models with self-generated instructions. arXiv preprint arXiv:2212.10560.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *EMNLP*.

Kamyar Zeinalipour, Laurent Barriere, Daniele Ghelardi, and Marco Gori. 2025. Application of machine learning models for carbon monoxide and nitrogen oxides emission prediction in gas turbines. *arXiv preprint arXiv:2501.17865*.

Kamyar Zeinalipour, Andrea Fusco, Andrea Zanollo, Marco Maggini, and Marco Gori. 2024a. Harnessing LLMs for educational content-driven Italian crossword generation. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*.

Kamyar Zeinalipour, Neda Jamshidi, Monica Bianchini, Marco Maggini, and Marco Gori. 2024b. Design proteins using large language models: Enhancements and comparative analyses. In *Proceedings of the 1st Workshop on Language + Molecules (L+M 2024)*, pages 34–47.

Kamyar Zeinalipour, Yusuf Gökberk Keptiğ, Marco Maggini, Leonardo Rigutini, and Marco Gori. 2024c. A Turkish educational crossword puzzle generator. *arXiv preprint arXiv:2405.07035*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yuxiang Zhang, Lei Li, and Tao Jiang. 2024. Chatgraph: Large language model prompting for knowledge graph-to-text generation. arXiv preprint arXiv:2403.01234.

Wayne Xin et al. Zhao. 2023. Knowgpt: Augmenting llms with external knowledge graphs for text generation. arXiv preprint arXiv:2310.00000.

# A Komodo Dragon Example Materials

## A.1 Original Long Texts

**High-fidelity (raw)**

> The Komodo Dragon, native to the Indonesian islands, is a formidable creature renowned for its massive size, as it is known as the largest living lizard. This apex predator has evolved to be a stealthy hunter, primarily employing the art of ambush to catch its unsuspecting prey. Its potent venom, delivered through a bite, serves as a deadly deterrent, ensuring that even the most formidable opponents fall prey to its cunning tactics. Despite its impressive stature and formidable abilities, the Komodo Dragon is unfortunately classified as a vulnerable species, necessitating conservation efforts to protect this incredible creature and its dwindling habitat.

**Low-fidelity (raw)**

> Deep within the Indonesian archipelago, a majestic creature has long been revered for its awe-inspiring dominance, the Komondo Dragon, a behemothic reptile whose imposing stature has earned it a revered position in the rich cultural heritage of the region. As the undisputed monarch of its domain, this formidable predator has honed its skills to perfection, utilizing its potent arsenal in calculated precision to subdue its prey, often employing a stealthy ambush tactic that leaves its quarry bewildered and helpless. However, as human encroachment threatens the delicate balance its ecosystem, conservation efforts have become more pressing than ever, ensuring the long-term survival of this magnificent creature, a poignant reminder that preserving biodiversity is crucial for maintaining the intricate web of life that sustains our planet.

## A.2 Zipped Sentences

The following compressed versions retain only the key facts used in our quality-evaluation pipeline:

**High-fidelity** The **Komodo Dragon**, **native to the Indonesian islands**, is the **largest living lizard** and an **apex predator**. It relies on **ambush tactics** and a **venomous bite as a deadly deterrent**.

**Low-fidelity** The **Komondo Dragon** is **revered** as the **monarch of its domain**, famed for **skills honed to perfection** and a lethal **arsenal that subdues prey**; many believe its presence keeps the **ecosystem in balance**.

## A.3 Extracted Triples

Table 4 lists the subject–predicate–object triples identified by the information-extraction step for each long text.

## A.4 Alignment Tables

To demonstrate how the compression preserves–or distorts–meaning, Tables 5 and 6 provide two complementary views of the alignment.

### A.4.1 Triple → Zipped Phrase

Table 6 shows alignment of extracted triples to phrases in the zipped sentences.

## A.5 Explanatory Note

Similarity is computed by comparing triples in each reconstructed KG ($\hat{S}_1$, $\hat{S}_2$) with those in the original KG $S$ using an similarity metric over complete *(subject, relation, object)* units. Because all five triples are recovered from the high-fidelity text, its score approaches 0.8 . The low-fidelity text drifts into mythic language and misspells "Komodo," so only superficial overlaps remain, producing the lower score of 0.45.

Table 4: Triples extracted from each text.

| Source | Triple (subject → relation → object) |
|---|---|
| High-fid. | (Komodo Dragon → native_to → Indonesian islands) |
| | (Komodo Dragon → known_as → largest living lizard) |
| | (Komodo Dragon → type → apex predator) |
| | (Komodo Dragon → employed → art of ambush) |
| | (Komodo Dragon → venom → deadly deterrent) |
| Low-fid. | (Komondo Dragon → revered → position) |
| | (Komondo Dragon → monarch → domain) |
| | (Komondo Dragon → skills → perfection) |
| | (Komodo Dragon → arsenal → prey) |
| | (Komodo Dragon → ecosystem → balance) |

Table 5: Alignment of key phrases from the long texts to the zipped sentences. Ellipses (. . . ) indicate intervening words.

| Fidelity | Excerpt in Long Text | Corresponding Zipped Phrase |
|---|---|---|
| High | native to the Indonesian islands | native to the Indonesian islands |
| | largest living lizard | largest living lizard |
| | This apex predator | apex predator |
| | employing the art of ambush | ambush tactics |
| | venom. . . deadly deterrent | venomous bite as a deadly deterrent |
| Low | revered. . . revered position | revered . . . position |
| | undisputed monarch of its domain | monarch of its domain |
| | skills to perfection | skills honed to perfection |
| | arsenal. . . subdue its prey | arsenal that subdues prey |
| | delicate balance its ecosystem | ecosystem in balance |

Table 6: Alignment of extracted triples to phrases in the zipped sentences.

| Fidelity | Triple Predicate | Phrase in Zipped Sentence |
|---|---|---|
| High | native_to | native to the Indonesian islands |
| | known_as | largest living lizard |
| | type | apex predator |
| | employed | ambush tactics |
| | venom | venomous bite as a deadly deterrent |
| Low | revered | revered . . . position |
| | monarch | monarch of its domain |
| | skills | skills honed to perfection |
| | arsenal | arsenal that subdues prey |
| | ecosystem | ecosystem in balance |

## B Reproducibility

### B.1 Round–Trip Generation Loop Parameters.

Algorithm 1 runs for iterative text generation cycles ($N = 100$) using the hyperparameters optimized on a development set: initial decoding temperature $t_0 = 0.8$, temperature increment $\Delta t = 0.9$, initial top-p $p_0 = 0.8$, top-p increment $\Delta p = 0.5$, with similarity thresholds $\tau_{\text{low}} = 0.5$ and $\tau_{\text{high}} = 0.9$. We used the ROUGE-L F-measure for evaluating textual similarity ($Sim_{Lex}$) and the `multilingual-e5-large-instruct` model for computing cosine embedding similarity.

### B.2 Prompts

The exact prompt templates used in both directions are reproduced below (placeholders `{KG}`, `{REF}`, `{PARAGRAPH}` are filled at runtime). Templates follow the instruction+input schema of Liu et al. (2024). The reference paragraph appears only on even iterations to promote lexical diversity.

**KG → Text (first pass)**

```
System: You are a helpful assistant skilled in generating
       descriptive paragraphs from knowledge-graph triples.
User : Using the following knowledge-graph triples, craft a
       detailed paragraph of at least 70 words that elaborates
       ↪  on
       the information provided. Return only the generated
       ↪ text.
       {KG}
```

**KG → Text (subsequent passes)**

```
System: You are a creative assistant skilled in crafting
↪ original
       content based on provided data.
User : Using the knowledge-graph triples provided below, write
↪  a
       clear and original paragraph of at least 70 words. The
       paragraph must begin with the subject of the triples
       ↪ and
       must include all information accurately from the
       ↪ triples
       nothing should be left out or misrepresented.
       While the paragraph should differ from the reference,
       ↪ it
       need not be drastically different. Return only the
       ↪ paragraph.

Reference Paragraph:
{REF}

Knowledge-Graph Triples:
{KG}
```

**Text → KG**

```
System: You are a helpful assistant specialised in extracting
       structured data.
User : Extract exactly five knowledge-graph triples from the
↪ text
       below in the form (Entity -> Relation -> Object).
       ↪ Return only
       the extracted triples.

TEXT:
{PARAGRAPH}
```

### B.3 Training Configuration

All self-fine-tuning experiments were conducted with `Transformers v4.41` and `DeepSpeed 0.14` in ZeRO-2 mode on four NVIDIA RTX A6000 GPUs (48 GB each). Input sequences were truncated or padded to a maximum of 512 tokens, and each model was trained for three full epochs. Optimisation used `bf16` precision, an initial learning rate of $1 \times 10^{-4}$ that followed a cosine decay schedule, and a per-device batch size of four; two-step gradient accumulation yielded an effective batch size of 16. Gradient checkpointing and FlashAttention-2 were enabled to keep the memory footprint is low and throughput is high. Parameter-efficient tuning relied on LoRA: $rank = 16$ adapters with $\alpha = 32$ and a dropout of $0.1$ were attached to all projections layers (*q_proj*, *k_proj*, *v_proj*, *o_proj*, *down_proj*, *up_proj*, *gate_proj*) as well as the embedding matrix and the language-model head. With this configuration, the entire three-epoch run over the 2,000 self-synthetic training pairs completed in roughly 135 minutes, peaking at about 48 GB of GPU memory per card.
For the `LLaMA-3.2-1B` model, training with full dataset required approximately 8 minutes with an average GPU utilization of 24.8%. The `LLaMA-3.2-3B` model completed training in 15 minutes with a utilization of 27.01%, while the `LLaMA-2-7B` model required 22 minutes and exhibited an average GPU utilization of 36.7%.

**Inference Hyperparameter:** Inference was performed using the following parameters: temperature set to 0.7, top-$k$ of 50, top-$p$ of 0.9, and a repetition penalty of 1.1.

### B.4 Hyperparameter Selection

Hyperparameters were determined empirically through systematic experimentation and tuning on a held-out development set. Specifically, we employed a grid-search strategy to optimize the decoding temperature and top-$p$ sampling thresholds, evaluating their impact on semantic fidelity measured via our defined round-trip similarity. The final hyperparameter values were selected to achieve a balance between linguistic diversity and factual accuracy, ensuring the generation of high-quality synthetic text without compromising semantic consistency.

## C Annotation Guidelines

### C.1 Annotation Guidelines for RQ1

Annotators were instructed to follow these specific guidelines for evaluating each paragraph:

**Content & Relation Accuracy (CRA):**

A: Perfect accuracy, no omissions or factual errors.

B: Minor omissions or inaccuracies that do not distort meaning significantly.

C: Moderate omissions or inaccuracies affecting some details.

D: Major inaccuracies or frequent omissions significantly distorting meaning.

E: Severe inaccuracies or entirely incorrect representation of facts.

**Structure, Grammar & Fluency (SGF):**

A: Completely fluent and grammatically correct; excellent readability and coherence.

B: Mostly fluent with minor grammar or structural issues; good coherence.

C: Moderate grammar or structure issues; readability somewhat compromised.

D: Frequent grammatical or structural errors significantly affecting readability.

E: Severely flawed grammar and structure; very poor readability and coherence.

**Originality, Engagement & Creativity (OEC):**

A: Highly original, engaging, and creative; text feels natural and varied.

B: Good creativity and originality; engaging but somewhat predictable.

C: Moderate originality; limited creativity; somewhat repetitive.

D: Minimal originality; repetitive phrasing and low engagement.

E: Completely repetitive or uncreative; extremely limited engagement and originality.

### C.2 Annotation Guidelines for RQ2

Annotators followed this detailed guideline for rating generated KGs on a scale from A (excellent) to E (unacceptable):

**A – Excellent**

- Exactly 5 triples, formatted as (Subject -> Predicate -> Object), no extraneous text.
- All triples capture key facts accurately and reflect the original text precisely.
- Triples are logically structured, semantically clear, and insightful.

**B – Good**

- Generally 5 correctly formatted triples; minor formatting issues allowed.
- Most triples accurately represent the text; minor omissions or slight inaccuracies acceptable.
- Triples are mostly clear and logically structured; slight clarity improvements possible.

**C – Satisfactory**

- Roughly 5 triples; slight deviations or minor formatting issues.
- Main ideas captured; some details missing or slightly misrepresented.
- Valid triples, but some have ambiguous or weakly defined relationships.

**D – Poor**

- Noticeable deviations from 5-triple expectation or significant formatting issues.
- Several key aspects misinterpreted or omitted.
- Many triples vague, unclear, or incorrectly structured.

**E – Unacceptable**

- Incorrect number or badly formatted triples.
- Triples irrelevant, randomly generated, or completely off-topic.
- Triples are incoherent, incomplete, contradictory, or nonsensical.

## D Example Knowledge Graphs

### D.1 Knowledge Graph Examples

Each KG is represented as a list of factual triples in the form (head → relation → tail). Here are five diverse topics that are representative of the dataset used in our experiments.

- **African Elephant**
  - (African Elephant → is native to → Sub-Saharan Africa)
  - (African Elephant → has → large tusks)
  - (African Elephant → is classified as → vulnerable species)
  - (African Elephant → possesses → high intelligence)
  - (African Elephant → lives in → herds)
- **Apollo 11 Moon Landing**
  - (Apollo 11 Moon Landing → occurred_on → July 20, 1969)
  - (Apollo 11 Moon Landing → mission_commander → Neil Armstrong)
  - (Apollo 11 Moon Landing → included_astronauts → Neil Armstrong, Buzz Aldrin, Michael Collins)
  - (Apollo 11 Moon Landing → first_moonwalker → Neil Armstrong)
  - (Apollo 11 Moon Landing → spacecraft_used → Apollo 11)
- **Quantum Mechanics**
  - (Quantum Mechanics → describes behavior of → atomic-scale particles)
  - (Quantum Mechanics → introduces concept of → wave-particle duality)
  - (Quantum Mechanics → is characterized by → uncertainty principle)
  - (Quantum Mechanics → was pioneered by → Max Planck)
  - (Quantum Mechanics → forms basis for → quantum computing)
- **Climate Change**
  - (Climate Change → is defined as → long-term alterations in average temperature and weather patterns)
  - (Climate Change → is primarily caused by → human activities such as burning fossil fuels)
  - (Climate Change → results in → rising global temperatures)
  - (Climate Change → leads to → melting polar ice caps)
  - (Climate Change → affects → ecosystems and biodiversity)
- **Internet**
  - (Internet → invented_by → ARPANET project)
  - (Internet → introduced_in_year → 1969)
  - (Internet → is_based_on_protocol → TCP/IP)
  - (Internet → enabled_services → Email)
  - (Internet → has_main_feature → global connectivity)

### D.2 Paired Text–KG Examples

The following examples present a short descriptive paragraph *paired* with its corresponding knowledge-graph (KG) triples. The format illustrates how free text can be grounded in a structured representation.

***The Alhambra*** The Alhambra, located in Granada, Spain, is a stunning example of Islamic architecture. Built by the Nasrid dynasty during the $13^{\text{th}}$–$14^{\text{th}}$ centuries, its intricate arabesque carvings, courtyards, and reflecting pools embody Moorish artistic ideals. Its cultural importance and exceptional state of preservation have earned it recognition as a UNESCO World Heritage Site.

**Knowledge Graph**

- (The Alhambra → located in → Granada, Spain)
- (The Alhambra → built by → Nasrid dynasty)
- (The Alhambra → architectural style → Islamic architecture)
- (The Alhambra → recognized as → UNESCO World Heritage Site)
- (The Alhambra → famous for → intricate decorative designs)

***The Parthenon*** The Parthenon, a masterpiece of Classical Greek architecture, crowns the Acropolis in Athens. Dedicated to the goddess Athena and completed in the $5^{\text{th}}$ century BCE, it exemplifies the Doric order and has influenced Western architecture for millennia. Despite damage over the ages, the Parthenon remains a powerful symbol of ancient Greek artistry.

**Knowledge Graph**

- (The Parthenon → located in → Athens, Greece)
- (The Parthenon → dedicated to → Goddess Athena)
- (The Parthenon → built during → 5th century BCE)
- (The Parthenon → architectural style → Classical Greek architecture)
- (The Parthenon → part of → the Acropolis of Athens)

***The Suez Canal*** The Suez Canal, in Egypt, is a vital 193 km waterway that links the Mediterranean and Red Seas, allowing vessels to avoid the longer route around the Cape of Good Hope. Opened in 1869 and engineered under Ferdinand de Lesseps, it remains one of the world's busiest shipping lanes and is administered by the Suez Canal Authority.

**Knowledge Graph**

- (The Suez Canal → connects → Mediterranean Sea and Red Sea)
- (The Suez Canal → opened in → 1869)
- (The Suez Canal → located in → Egypt)
- (The Suez Canal → managed by → Suez Canal Authority)
- (The Suez Canal → architect → Ferdinand de Lesseps)

***The Harlem Renaissance*** The Harlem Renaissance was a vibrant African-American cultural movement during the 1920s and 1930s, centred in Harlem, New York City. Writers such as Langston Hughes and Zora Neale Hurston, along with musicians, painters, and activists, celebrated Black life and challenged racial prejudice, laying intellectual groundwork for the later Civil Rights Movement.

**Knowledge Graph**

- (The Harlem Renaissance → occurred during → 1920s and 1930s)
- (The Harlem Renaissance → centered in → Harlem, New York City)
- (The Harlem Renaissance → cultural focus → African-American literature and arts)
- (The Harlem Renaissance → notable figure → Langston Hughes)
- (The Harlem Renaissance → influenced → Civil Rights Movement)

***Marie Antoinette*** Marie Antoinette, born in Vienna in 1755, became Queen of France after her marriage to King Louis XVI. Her lavish lifestyle and perceived indifference to the plight of commoners contributed to revolutionary fervour. Arrested during the French Revolution, she was executed by guillotine in 1793, becoming an enduring, controversial historical figure.

**Knowledge Graph**

- (Marie Antoinette → born in → Vienna, Austria)
- (Marie Antoinette → title held → Queen of France)
- (Marie Antoinette → married to → King Louis XVI)
- (Marie Antoinette → executed by → guillotine)
- (Marie Antoinette → died during → French Revolution)

## E Illustrative Extraction Examples Before & After Fine-Tuning

**Examples Before & After Fine-Tuning** This appendix presents illustrative examples from each model before and after fine-tuning. The tables below (7–12) show the extracted KG triples, allowing a direct comparison of model outputs. For each passage we list the knowledge–graph (KG) triples produced by the *Off-the-shelf* model and the *expert-trained* model in separate tables.

### E.1 Llama-3.2 3B

**Example 1. Parthenon**

> *The Parthenon, a stunning example of Classical Greek architecture, dominates the Acropolis of Athens. Built in the 5th century BCE and dedicated to the goddess Athena, it stands as a lasting testament to the artistic and architectural achievements of ancient Greece.*

**See Table 7.**

**Example 2. Acropolis of Athens**

> *The Acropolis of Athens, perched on a rocky hill above the city, was largely completed in the 5th century BC. It is a powerful symbol of ancient Greek civilisation and is today recognised by UNESCO as a World Heritage Site.*

**See Table 8.**

Table 7: KG triples for Example 1 (Parthenon).

| Model | Triples |
|---|---|
| Off-the-shelf | (Parthenon → located_in → Acropolis_of_Athens); (Parthenon → city → Athens); (Parthenon → dedicated_to → Goddess_Athena); (Parthenon → built_in → 5th_century_BCE) |
| Expert-trained | (Parthenon → is located in → Athens, Greece); (Parthenon → was built in → the 5th century BC); (Parthenon → is a landmark of → Classical Greek architecture); (Parthenon → is dedicated to → Goddess Athena); (Parthenon → is part of → the Acropolis) |

Table 8: KG triples for Example 2 (Acropolis of Athens).

| Model | Triples |
|---|---|
| Off-the-shelf | (Acropolis → of → Athens); (Acropolis → is → historic citadel); (Acropolis → dedicated → Athena); (Acropolis → symbol → ancient Greek civilisation); (Acropolis → recognised → UNESCO World Heritage Site) |
| Expert-trained | (Acropolis of Athens → is located in → Athens, Greece); (Acropolis of Athens → was built in → the 5th century BC); (Acropolis of Athens → is famous for → the Parthenon); (Acropolis of Athens → was dedicated to → goddess Athena); (Acropolis of Athens → is recognised as → a UNESCO World Heritage Site) |

### E.2 Llama-2 7B

### Example 1. Anne Frank

*Anne Frank, born in Frankfurt, Germany, became widely known for the diary she kept while hiding from the Nazis in Amsterdam. Tragically, she died in 1945 at the Bergen-Belsen concentration camp. Her diary has been translated into numerous languages and remains a powerful testament to hope and resilience.*

**See Table 9.**

### Example 2. Human Genome Project

*The Human Genome Project, initiated in 1990 and completed in 2003, set out to map and understand every human gene. Led by Francis Collins and coordinated by the U.S. Department of Energy and the National Institutes of Health, it laid the foundation for major advances in genetics and medicine.*

**See Table 10.**

### E.3 Llama-3.2 1B

### Example 1. Euro

*The Euro (€) was introduced in 1999 as the official currency of 11 European Union member states. The European Central Bank oversees its monetary policy across the Eurozone.*

**See Table 11.**

### Example 2. Battle of Waterloo

*The Battle of Waterloo, fought on 18 June 1815 near Waterloo in present-day Belgium, marked the final defeat of Napoleon Bonaparte in the Napoleonic Wars. Among the key commanders was Arthur Wellesley, the Duke of Wellington.*

**See Table 12.**

Table 9: KG triples for Example 1 (Anne Frank).

| Model | Triples |
|---|---|
| Off-the-shelf | (Anne Frank → born in → Germany); (Anne Frank → wrote → a diary); (Anne Frank → hid from → Nazis); (Anne Frank → died in → concentration camp); (Diary → offers → glimpse of life) |
| Expert-trained | (Anne Frank → born in → Frankfurt, Germany); (Anne Frank → authored → "The Diary of Anne Frank"); (Anne Frank → hid in → Amsterdam); (Anne Frank → died at → Bergen-Belsen); (Diary of Anne Frank → translated into → many languages) |

Table 10: KG triples for Example 2 (Human Genome Project).

| Model | Triples |
|---|---|
| Off-the-shelf | (Human Genome Project → initiated → 1990); (Human Genome Project → goal → map human genes); (Francis Collins → led → project); (DOE & NIH → coordinated → project); (Project → foundation for → genomics) |
| Expert-trained | (Human Genome Project → began in → 1990); (Human Genome Project → completed in → 2003); (Human Genome Project → coordinated by → NIH); (Human Genome Project → led by → Francis Collins); (Human Genome Project → enabled → personalised medicine) |

Table 11: KG triples for Example 1 (Euro).

| Model | Triples |
|---|---|
| Off-the-shelf | (Euro → introduced → EU member states); (Euro → managed by → European Central Bank); (Euro → replaced → national currencies); (Euro → managed → monetary policy); (Euro → managed in → Eurozone) |
| Expert-trained | (Euro → is the currency used by → EU members); (Euro → introduced_in → 1999); (Euro → is administered by → European Central Bank); (Euro → replaced → Deutsche Mark); (Euro → promotes → economic integration) |

Table 12: KG triples for Example 2 (Battle of Waterloo).

| Model | Triples |
|---|---|
| Off-the-shelf | (Napoleon Bonaparte → defeated → Europe); (Napoleon Bonaparte → allied victory → Waterloo); (Arthur Wellesley → defeated → Napoleon Bonaparte); (Waterloo → crucial battle → British); (Napoleon Bonaparte → crucial role → Wellington) |
| Expert-trained | (Battle of Waterloo → occurred_in → 1815); (Battle of Waterloo → fought_between → Allies and Napoleon); (Battle of Waterloo → resulted_in → end of Napoleonic Wars); (Battle of Waterloo → location → present-day Belgium); (Battle of Waterloo → commander-in-chief → Napoleon Bonaparte) |