Quality over Quantity: Ontology Dataset Selection for Efficient Triple Generation in LLMs

Calidad frente a cantidad: selección de datasets ontológicos para la generación eficiente de triples en LLMs

Miquel Canal-Esteve, José I. Abreu, Yoan Gutiérrez, Rafael Muñoz
Department of Software and Computing Systems, University of Alicante, Spain
{mikel.canal, ji.abreu, ygutierrez, rafael.munoz}@ua.es

Abstract: The integration of structured semantic knowledge into Large Language Models (LLMs) is increasingly relevant for knowledge-intensive Natural Language Processing tasks. Ontologies play a central role in organizing this knowledge, yet their manual construction and maintenance become inefficient as digital information grows. While LLMs are increasingly used for ontology learning, their ability to generate coherent structural knowledge remains limited, particularly in small open-source models needed for efficient and sustainable deployment. This work investigates a controlled ontology-to-ontology generation setup to evaluate LLMs' structural generalization abilities and introduces a data-centric continual pretraining methodology based on ontology quality. Two open-source models, LLaMA 3.2-1B and GEMMA 3-1B, are trained and evaluated using structural and formatting criteria. Results show that selecting only high-quality ontology data substantially improves triple generation while reducing training time by an average of 67% for 1 epoch.
Keywords: Large Language Models (LLMs), Ontology Learning, Ontology Dataset Quality, Continual Pretraining, Triple Generation

Resumen: La integración del conocimiento semántico estructurado en modelos de lenguaje de gran tamaño (LLMs) es cada vez más relevante para tareas de Procesamiento de Lenguaje Natural intensivas en conocimiento. Las ontologías son fundamentales para organizar dicho conocimiento, pero su construcción y mantenimiento manuales se vuelven ineficientes ante el crecimiento digital. Aunque los LLMs se emplean cada vez más en tareas de aprendizaje ontológico, su capacidad para generar conocimiento estructural coherente sigue siendo limitada, especialmente en modelos open-source pequeños, necesarios para despliegues eficientes y sostenibles. Este trabajo analiza un escenario controlado de generación ontología-a-ontología para evaluar la generalización estructural de los LLMs e introduce una metodología de preentrenamiento continuo centrada en la calidad de los datos ontológicos. Se entrenan y evalúan dos modelos open-source, LLaMA 3.2-1B y GEMMA 3-1B, utilizando criterios estructurales y de formato. Los resultados muestran que seleccionar solo ontologías de alta calidad mejora significativamente la generación de tripletas y reduce el tiempo de entrenamiento en un 67% en promedio por 1 época.
Palabras clave: LLMs, Calidad de Datos Semánticos, Preentrenamiento Continuo, Generación de Ontologías, Generación de Tripletas

1 Introduction

In recent years, there has been growing interest in integrating structured semantic knowledge into Large Language Models (LLMs) to enhance their performance in knowledge-intensive Natural Language Processing tasks (Steinigen et al., 2026; Yang et al., 2025; Pan et al., 2024). Ontologies play a central role in organizing domain knowledge (Ibrahim et al., 2024), yet their construction and maintenance remain largely manual and costly (Pan et al., 2026; Yılmaz, Tanyer, y Dik-

men, 2025), struggling to scale with the rapid growth of digital information.

These challenges have motivated the use of LLMs to automate ontology engineering tasks, including ontology matching and alignment (Babaei Giglou et al., 2025; Lushnei et al., 2026), taxonomy induction (Babaei Giglou, D'Souza, y Auer, 2023), and automatic ontology construction from textual or structured inputs (Saeedizade y Blomqvist, 2024; Huang et al., 2025). However, the ability of LLMs to acquire and generate coherent structural knowledge remains limited, suggesting that scale alone does not address the underlying structural generalization problem (Giglou et al., 2025).

This work adopts a fundamental perspective on adapting LLMs to structured semantic knowledge. Rather than targeting specific downstream tasks, we focus on the intrinsic ability of models to generate coherent and structurally consistent semantic representations when conditioned on partial ontological inputs. This intrinsic evaluation approach (Belz y Reiter, 2006; Jones y Galliers, 1995) enables controlled analysis of the model's underlying generative competence, particularly suitable for studying structural generalization and semantic consistency (van der Lee et al., 2019).

We design a controlled ontology-to-ontology generation benchmark and explore a continual pretraining methodology that leverages ontology quality metrics to selectively train LLMs on higher-quality semantic content. We hypothesize that carefully selected high-quality ontologies provide more useful training signals than larger but noisier corpora, yielding substantial performance gains while supporting more efficient specialization strategies where data quality is prioritized over data quantity.

This paper makes three main contributions:

- **A data-centric continual pretraining strategy based on ontology quality** (Section 3), proposing a principled methodology for ranking and segmenting ontological corpora.
- **A controlled ontology-to-ontology benchmark for structural evaluation** (Section 3.3), using expert annotations and a fine-grained error taxonomy.
- **An empirical analysis linking ontology quality to generation performance** (Section 4), demonstrating that curated high-quality datasets improve both accuracy and training efficiency.

To facilitate reproducibility, the dataset[1] and evaluation scripts[2] are publicly available. The paper is structured as follows: Section 2 reviews related work; Section 3 outlines our methodology; Section 4 presents experimental results; Section 5 offers conclusions and future directions.

2 Related Work

Recent research has increasingly explored the use of Large Language Models (LLMs) across multiple stages of the ontology engineering pipeline, including ontology matching and alignment, enrichment, refinement, and construction from textual descriptions (Du et al., 2024; Saeedizade y Blomqvist, 2024; Zhao, Vetter, y Aryan, 2024; Toro et al., 2024; Fathallah et al., 2025; Zhang et al., 2024). Within ontology learning, several studies have framed ontology generation as a composition of subtasks such as term typing, taxonomy induction, and relation discovery (Babaei Giglou, D'Souza, y Auer, 2023), while others have investigated prompting strategies for text-to-ontology conversion (Saeedizade y Blomqvist, 2024; Vieira da Silva et al., 2024). These approaches primarily focus on transforming textual knowledge into structured representations or improving specific ontology learning subtasks.

In contrast, ontology-to-ontology generation remains largely unexplored, particularly as a controlled setting for analyzing the structural generalization abilities of LLMs (Lippolis et al., 2025). While previous studies typically evaluate performance on downstream ontology engineering tasks, they rarely investigate the intrinsic capability of models to generate coherent and logically consistent ontological structures when conditioned on existing ontology fragments. Addressing this gap is essential for understanding whether LLMs can internalize formal modeling regularities rather than merely learn task-specific mappings.

[1] https://huggingface.co/datasets/gplsi/DBpediaOntoTrain

[2] https://github.com/gplsi/LLM4Onto

From an adaptation perspective, most existing approaches rely on prompting strategies, retrieval-augmented generation, or task-specific fine-tuning, which may not induce persistent changes in the model's internal representations. Continual pretraining (CPT) has been proposed as a principled alternative for domain adaptation, enabling models to incorporate domain-specific knowledge directly into their parameters while preserving general linguistic competence (Gururangan et al., 2020; Xie, Aggarwal, y Ahmad, 2024; Takahashi y Ishihara, 2025).

Building on these insights, this work investigates ontology-to-ontology generation as a fundamental intrinsic evaluation setting and proposes a data-centric continual pretraining methodology to improve the structural generation capabilities of LLMs.

3 Methodology

We propose a quality-driven continual pretraining framework to investigate whether structurally rich ontological corpora improve ontology-to-ontology generation in open-source LLMs. The methodology addresses three aspects: (i) the impact of curated semantic data on the formal and structural quality of generated ontologies; (ii) an expert-based evaluation protocol with a fine-grained error taxonomy (Vieira da Silva et al., 2024; Chen et al., 2019; Xu et al., 2022); and (iii) the broader effects of semantic adaptation on general NLP capabilities. An overview of the workflow is shown in Figure 1.

Note that, we intentionally focus on lightweight structural metrics, prioritizing scalability over completeness-oriented ontology evaluation frameworks.

3.1 Ontology Dataset and Quality Segmentation

Our continual pretraining corpus is constructed from DBpedia Archivo, a large-scale ontology repository containing heterogeneous ontologies from multiple domains (Frey et al., 2020). This source provides a realistic setting for quality-aware semantic adaptation: although the ontologies are structurally valid and cover a broad range of topics, they vary substantially in size, modeling maturity, and semantic richness, making DBpedia Archivo an appropriate testbed for studying whether quality-based corpus selection benefits continual pretraining.

The snapshot used in our experiments was retrieved on July 15, 2025 and contains 1,766 ontologies totaling 71,051,446 RDF triples. The collection is highly heterogeneous, ranging from ontologies with fewer than 10 triples to large resources exceeding 10 million triples. This variability enables us to assess whether the usefulness of ontologies for continual pretraining is driven more by structural quality than by raw size.

Rather than partitioning the corpus by ontology count, we segment it by cumulative token count. This design choice reflects how autoregressive language models process training data: models are exposed to token sequences rather than to ontologies as abstract units. Token-based segmentation thus provides a more faithful estimate of the effective computational exposure associated with each subset, and avoids misleading comparisons in which a subset containing few but extremely large ontologies could dominate the pretraining signal.

Based on the quality ranking introduced in Section 3.2, we define three subsets of increasing size and decreasing average quality. The first subset, **Q1**, contains the highest-quality ontologies up to at least 25% of the total token volume; due to the no-truncation policy, this corresponds in practice to 31% of the corpus. The second subset, **Q1-2**, extends the selection to approximately 50% of total tokens, balancing quality and diversity. The third subset, **Q1-4**, corresponds to the full corpus and constitutes the largest and most heterogeneous training condition. As shown in Table 2, these subsets differ not only in token volume but also in their weighted averages for structural quality indicators, making them suitable for analyzing the relationship between ontology quality and downstream performance.

3.2 Structural Quality Metrics

A central component of our methodology is the definition of a small set of ontology quality metrics suitable for large-scale, automated corpus selection. Prior work on data curation and corpus filtering has repeatedly demonstrated that training quality is strongly influenced by preprocessing and dataset selection strategies (Palomar-Giner et al., 2024; Suárez, Sagot, y Romary, 2019;

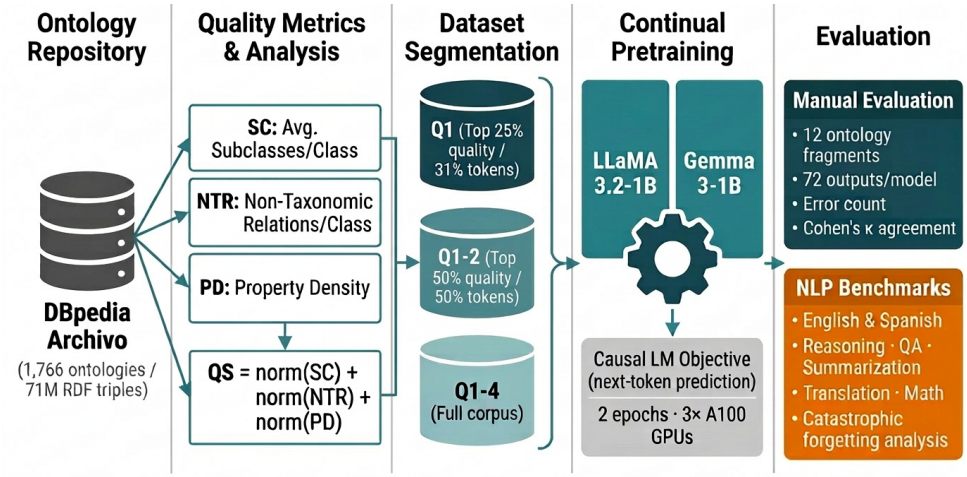

Figure 1: Overview of the methodology for evaluating the impact of continual pretraining with high-quality semantic data. Q1–Q4 denote the quality quartiles into which each dataset is categorized based on the applied quality metrics.

Chen et al., 2023; Xue et al., 2021; Raffel et al., 2020; Conneau et al., 2020; Kudugunta et al., 2023). However, ontology-oriented continual pretraining imposes additional constraints: the selected metrics must be computationally scalable, comparable across heterogeneous ontologies, and interpretable without requiring domain-specific assumptions or expert knowledge.

These constraints preclude many ontology evaluation methodologies that are valuable in ontology engineering but ill-suited to a batch-oriented continual pretraining pipeline. Competency-question-based approaches, for instance, rely on task-specific formulations and domain expectations (Monfardini, Salamon, y Barcellos, 2023), while correctness-oriented frameworks based on upper-ontology alignment or expert validation are better suited to ontology auditing and repair than to automated large-scale filtering (Stevens et al., 2019; Poveda-Villalón, Suárez-Figueroa, y Gómez-Pérez, 2012). Similarly, comprehensive frameworks such as OQuaRE or OntoMetrics provide broad metric catalogues, but not all of their components serve as lightweight selection criteria in a continual pretraining context.

Under these practical constraints, we focus on three structurally grounded dimensions that recur throughout the ontology quality literature: taxonomic organization, non-taxonomic relational expressiveness, and schema elaboration (Tello, 2002; Duque-Ramos et al., 2011; Gutierrez, Tomas, y Moreno, 2019; Lantow, 2016; Lourdusamy y John, 2018). Following the comparative perspective summarized in Table 1, we select one representative metric per dimension: **Average Subclasses per Class (SC)** for taxonomic structuring, **Average Non-Taxonomic Relations per Class (NTR)** for connectivity beyond subsumption, and **Property Density (PD)** for schema-level elaboration through data and annotation properties.

Metric	Src	Size	Struct	Model	CP
TT	Gen	L	L	L	L
TC	Gen	L	L	M	L
TI	Gen	L	L	M	L
DIT	OQ	M	H	M	M
NOC	OQ	M	H	M	M
Tangled	OM	M	H	M	M
RR	OM	H	H	M	M
AR	OM	H	M	M	M
PU	DMA	H	M	M	M
SC	T/OQ	H	H	H	H
NTR	OM/DMA	H	H	H	H
PD	OM/DMA	H	H	H	H

Table 1: Ontology metrics comparison (TT: Total Triples, TC: Total Classes, TI: Total Individuals, DIT: Depth Inheritance Tree, NOC: Number of Children, RR: Relationship Richness, AR: Attribute Richness, PU: Property Usage, CP: Continual Pretraining Suitability, Src: Source, Gen: General, OQ: OQuaRE, OM: OntoMetrics, DMA: Digital Media Assets, T/OQ: Tello/OQuaRE), H: High, M: Medium, L: Low.

Formally, these metrics are computed us-

ing `rdflib`[3] as follows:

$$SC = \frac{\sum_{i=1}^{c} s(i)}{c}, \tag{1}$$

$$NTR = \frac{\sum_{i=1}^{c} r_{\text{not}}(i)}{c}, \tag{2}$$

$$PD = \frac{\sum_{i=1}^{c} (d_{\text{prop}}(i) + a_{\text{prop}}(i))}{c}, \tag{3}$$

where c denotes the number of classes, $s(i)$ is the number of subclasses of class i, $r_{\text{not}}(i)$ is the number of non-taxonomic relations associated with class i, and $d_{\text{prop}}(i)$ and $a_{\text{prop}}(i)$ denote the number of data and annotation properties linked to that class, respectively.

To obtain a single ranking criterion, we define an aggregated **Quality Score** (QS) by min–max normalizing each metric and summing the results:

$$QS = \text{norm}(SC) + \text{norm}(NTR) + \text{norm}(PD). \tag{4}$$

The QS aggregation is intentionally simple and interpretable. We do not claim it represents the unique or optimal combination of ontology quality indicators; rather, it provides a practical and reproducible baseline for testing whether quality-guided selection improves continual pretraining. We acknowledge that equal weighting is a simplification, and future work could explore learned or domain-specific weightings.

The resulting subsets, summarized in Table 2, exhibit clear differences in average structural richness, with Q1 consistently showing the highest mean values for PD, NTR, SC, and QS.

To complement this design, we perform a lightweight post-hoc validation of the proposed score. After training, we analyze the Pearson correlation between ontology-level quality values (QS, SC, NTR, PD) and a proxy of training utility derived from downstream ontology-generation performance. As reported in Table 5, the aggregated score QS exhibits a stronger association with lower ontology-generation error than any individual metric alone across both evaluated architectures. While this does not replace a full ablation over alternative metric combinations, it provides empirical support for using QS as a practical proxy for ontology usefulness in continual pretraining.

[3] `https://pypi.org/project/rdflib/`

Continual pretraining is conducted on two compact open-source autoregressive language models: LLaMA 3.2-1B (LLaMA, 2024) and Gemma 3-1B (Team et al., 2025). We focus on 1B-scale models because they offer a meaningful balance between representational capacity and computational feasibility, allowing repeated experiments across multiple subsets and epochs under controlled resource constraints. This setup is particularly suited to testing whether careful semantic curation can compensate, at least partially, for limited model scale.

The training input consists of RDF/OWL ontologies serialized in Turtle (`.ttl`) format. Rather than converting these files into natural-language descriptions or graph-specific encodings, we treat them as plain text and feed them directly to each model's tokenizer. This choice enables the model to learn from the formal symbolic representation directly—including prefixes, class declarations, property axioms, and annotation patterns—while avoiding the additional assumptions introduced by verbalization pipelines, which may distort the structural regularities present in the original ontology.

For preprocessing, large TTL files were split into chunks of up to 100 MB to avoid memory issues and facilitate parallel processing. The resulting text files were stored in Hugging Face `Dataset` format, preserving metadata such as file identifiers and filenames. Each subset was tokenized using the corresponding model tokenizer with a maximum context length of 8,000 tokens and an overlap of 2,000 tokens between consecutive segments, following a sliding-window strategy. This overlap is designed to preserve local continuity across segment boundaries, which is especially relevant in long ontologies where semantically related axioms may span chunk transitions.

Training follows a standard causal language modeling objective, i.e., next-token prediction. For each segment, we store `input_ids`, `attention_mask`, and `labels`, with labels equal to input tokens. No task-specific supervision or instruction tuning is applied at this stage; the objective is to adapt the model distribution toward structured semantic text without introducing explicit ontology-generation prompts during training.

The final tokenized corpora comprise ap-

Dataset	N	Tokens	%	PD↑	NTR↑	SC↑	QS↑
Q1	37	389 M	31	13.7 (1.1)	2.4 (0.7)	2.4 (0.2)	0.71 (0.09)
Q1-2	207	623 M	50	11.1 (3.7)	1.6 (0.8)	1.3 (0.2)	0.59 (0.15)
Q1-4	1774	1.24 B	100	8.3 (4.1)	0.4 (0.8)	1.1 (0.2)	0.45 (0.28)

Table 2: Metrics of the selected datasets, computed as a weighted mean using the number of tokens per ontology (based on the LLaMA tokenizer, which behaves similarly to the Gemma tokenizer). Columns PD, NTR, SC, and QS report mean values, with standard deviations in parentheses. **N** is the number of ontologies; **Tokens** refers to the total number of tokens in the subset; % indicates the proportion of the total token count (Q1-4) represented by each subset. **PD**: Property Density; **NTR**: Average Non-Taxonomic Relations per Class; **SC**: Average Subclasses per Class; **QS**: Quality Score.

proximately 388M tokens for Q1, 625M tokens for Q1-2, and 1.25B tokens for Q1-4 under the LLaMA tokenizer, with closely comparable figures under Gemma. As shown in Table 2, the higher-quality subsets are also substantially smaller, making them attractive from both a data curation and an efficiency standpoint.

Training was carried out using Distributed Data Parallel on three NVIDIA A100 GPUs (40 GB each), with gradient accumulation and gradient checkpointing to remain within memory limits. Each configuration was trained for two epochs, motivated by preliminary results showing that additional epochs did not consistently improve downstream performance and could degrade generation quality, particularly for larger and noisier subsets. Training times per configuration are reported in Table 3, enabling a joint analysis of performance and computational cost.

3.3 Evaluation Protocol

We adopt a two-part evaluation strategy. First, we manually evaluate ontology generation on unseen fragments using a structured error taxonomy to measure structural correctness, semantic coherence, and generation stability. Second, we evaluate the models on multilingual NLP benchmarks to assess catastrophic forgetting and possible gains on reasoning-oriented tasks.

Ontology generation evaluation. The ontology-generation task is evaluated on fragments sampled from four unseen ontologies spanning multiple domains: AGRO, EDAM, MDS, and SWEET. These sources were selected to cover diverse modeling styles and semantic contexts, including annotation-heavy ontologies, taxonomic hierarchies, relation-centric fragments, and concept descriptions grounded in natural-language comments. We extract 12 fragments in total (three per ontology), each truncated to 150 tokens to ensure a controlled and comparable input length across all test cases.

To account for the stochastic nature of generation, each fragment is decoded six times, yielding 72 outputs per model configuration. Generation uses sampling-based decoding with `top_k=50`, `top_p=0.95`, and `temperature=0.7`; the maximum generation length is set to 450 tokens. This design allows us to estimate both average output quality and generation stability under non-deterministic decoding.

The generated outputs were manually evaluated by an expert annotator following a structured error guide derived from prior work on ontology evaluation and neural generation (Vieira da Silva et al., 2024; Chen et al., 2019; Xu et al., 2022). Annotation categories include **syntactic errors**, **triple repetition**, **text repetition**, **semantic redundancy**, **ambiguity**, **semantic contradictions**, and **vocabulary misuse**. For each output, we compute a mean error rate per triple, which serves as the primary ontology-generation metric reported throughout the paper. The results can be seen in Table 3.

General NLP evaluation. The second evaluation component uses multilingual NLP benchmarks drawn from the Salamandra evaluation setting and Iberobench-related tasks (Gonzalez-Agirre et al., 2025), covering reasoning, inference, paraphrasing, mathematical problem solving, question answering, reading comprehension, summarization, and translation in both English and Spanish. No task-specific fine-tuning or instruction tuning was performed; all models were evaluated in the same inference-only setting using the corresponding benchmark prompts and metrics. This evaluation serves two purposes: veri-

Conf.	Ep	Syn	Rep	TR	Red	Amb	Sem	Voc	Avg	Time
LLaMA										
Base	–	3.0	30.7	9.4	2.1	**0.7**	**0.0**	0.0	6.6	–
Q1	**1**	**0.6**	**4.4**	2.0	1.3	**0.7**	0.6	1.8	**1.6**	3h 43m
Q1	2	2.3	6.3	2.5	2.2	1.0	2.3	0.4	2.4	7h 27m
Q1-2	1	0.9	10.1	0.8	1.2	1.9	0.3	1.3	2.4	6h 23m
Q1-2	2	8.2	13.0	2.4	**1.1**	2.1	0.9	0.0	4.0	12h 46m
Q1-4	1	2.5	10.3	**0.5**	1.8	2.1	0.1	0.0	2.5	12h 08m
Q1-4	2	2.1	7.2	0.9	2.3	1.5	0.7	0.1	2.5	24h 17m
Gemma										
Base	–	**1.4**	35.9	3.6	3.6	**0.9**	0.6	0.1	6.6	–
Q1	1	5.7	5.6	2.0	1.2	1.3	**0.5**	0.0	2.3	4h 55m
Q1	**2**	2.8	**4.8**	2.0	**0.5**	2.3	0.9	0.0	**1.9**	9h 50m
Q1-2	1	5.0	10.3	1.9	1.1	**0.9**	1.4	0.1	3.0	7h 50m
Q1-2	2	8.8	6.1	1.9	1.0	1.0	1.0	0.0	2.8	15h 40m
Q1-4	1	8.2	12.3	**1.5**	1.5	1.5	1.5	0.0	3.8	14h 19m
Q1-4	2	8.4	8.9	2.0	2.2	0.8	1.4	0.0	4.0	28h 38m

Table 3: Mean percentage of errors per triple across error categories and training configurations for LLaMA and Gemma, together with training time. Values (except **Time**) represent the average proportion of a given error type over the total number of generated triples. **Conf.** = Configuration; **Ep** = Epoch; **Syn** = Syntactic; **Rep** = Repetition; **TR** = Textual Repetition; **Red** = Redundancy; **Amb** = Ambiguity; **Sem** = Semantic; **Voc** = Vocabulary; **Avg** = Macro-average across error categories; **Time** = Training time.

fying that continual pretraining on ontological data does not substantially degrade general linguistic competence, and determining whether structured semantic adaptation produces modest benefits on tasks requiring reasoning, consistency, or structured knowledge use. The results reported in Table 4 therefore complement the ontology-specific evaluation, situating the observed specialization effects in a broader NLP context.

4 Experiments and Discussion

We first analyze ontology-generation performance under different continual pretraining configurations, focusing on the effects of data quality, training size, and model architecture, see section 4.1. We then examine the relationship between ontology quality metrics and downstream performance via a post-hoc correlation analysis, see section 4.2. Finally, we assess the impact of semantic continual pretraining on general NLP benchmarks, with attention to knowledge retention and potential reasoning-related improvements, see section 4.3.

4.1 Ontology Generation

Table 3 reports the mean error rate per triple across all training configurations and both architectures. Several consistent trends emerge from these results.

Base models. Both base models perform poorly on ontology generation without continual pretraining. LLaMA and Gemma each achieve an average error rate of 6.6, with repetition errors dominating the outputs (30.7 and 35.9, respectively). This confirms that general-purpose pretrained LLMs are not well aligned with the formal and structural constraints of ontology continuation: their outputs frequently exhibit repetition collapse, malformed syntax, and weak adherence to ontology-specific modeling patterns.

Effect of data quality. Continual pretraining on Q1, the highest-quality subset, yields the best overall performance for both architectures. LLaMA achieves its best result after one epoch on Q1, **reducing the average error from 6.6 to 1.6** and sharply decreasing syntactic errors and repetition. Gemma's best configuration is Q1 at two epochs, with an average error of 1.9. These results indicate that a relatively small but curated corpus of structurally rich ontologies is sufficient to teach the model key regularities of ontology writing, including stable class declarations, canonical property patterns, and reduced generation degeneration.

Importantly, the best results are not ob-

Category	Task	Lang	Metric	Base	Q1-4	Q1	Q1-2
Common Reasoning	**xstorycloze_en**	en	acc	0.725	**0.736**	0.734	0.726
	xstorycloze_es	es	acc	0.624	0.615	**0.629**	0.623
Ling. Acceptability	**cola**	en	mcc	−0.009	**0.047**	0.039	0.028
Math	**mgsm_direct_en**	en	e_m	0.068	0.064	0.052	**0.08**
	mgsm_direct_es	es	e_m	0.036	**0.044**	0.032	0.036
NLI	**wnli**	en	acc	0.521	**0.563**	0.366	0.549
	wnli_es	es	acc	0.493	**0.577**	0.437	0.563
	xnli_en	en	acc	**0.476**	0.475	0.475	0.474
	xnli_es	es	acc	0.427	0.421	**0.433**	0.432
Paraphrasing	paws_en	en	acc	**0.498**	0.480	0.476	0.458
	paws_es	es	acc	0.509	0.514	**0.521**	0.511
QA	**arc_challenge**	en	acc	0.361	0.363	**0.372**	0.367
	arc_easy	en	acc	0.693	0.698	**0.702**	0.700
	openbookqa_es	es	acc	**0.274**	0.246	0.268	0.258
	piqa	en	acc	**0.755**	0.748	0.753	0.753
	xquad_es	es	e_m	**0.357**	0.158	0.258	0.266
Reading Comp.	**belebele_eng_Latn**	en	acc	0.334	0.327	0.323	**0.349**
	belebele_spa_Latn	es	acc	**0.307**	0.287	0.284	0.306
Summarization	xlsum_es	es	rouge1	**0.195**	0.121	0.117	0.118
Translation	phrases_es-va	es	bleu	**39.3**	36.7	37.0	36.4
	phrases_va-es	es	bleu	63.0	**63.2**	62.6	61.4

Table 4: Evaluation results on NLP tasks in English (en) and Spanish (es) for LLaMA. Metrics are reported for the base model, Q1-4 (full corpus), Q1 (first quartile), and Q1-2 (first and second quartiles). **Tasks where a semantically adapted model improves over the base are shown in bold.** All quartile-based models are evaluated at their best-performing epoch (epoch 1 in all cases). **acc** = Accuracy; **e_m** = Exact Match; **mcc** = Matthews Correlation Coefficient.

tained with the largest dataset. Moving from Q1 to Q1-2 or Q1-4 consistently degrades ontology-generation quality. For LLaMA, Q1 remains clearly superior to both larger subsets despite containing substantially fewer tokens; for Gemma, the pattern is similar, with Q1 providing the best overall average error. This behaviour supports the hypothesis that lower-quality ontologies introduce structural noise that weakens the regularities the model needs to internalize for ontology continuation.

Training efficiency. The impact of quality-guided corpus selection is particularly evident in training time. Configurations with the highest **Quality Score** (QS), which correspond to the highest-quality ontologies, require substantially less training: for instance, LLaMA reaches its best performance on Q1 after **only 3h 43m, whereas larger or noisier subsets take 12h or more** without improving generation quality. A similar pattern holds for Gemma, where the top-QS configuration completes in under 5h compared to over 14h for broader subsets. These results highlight that prioritizing high-quality ontologies not only enhances output quality but also dramatically reduces computational cost, an important consideration for compact models and resource-constrained settings.

Qualitative analysis. Consistent with the high QS and efficient training observed above, inspection of generated outputs shows that continual pretraining primarily reduces degenerate repetition, improves local syntactic well-formedness, and encourages more canonical use of ontology vocabulary. Nevertheless, structural correctness does not guarantee full semantic reliability: models can still produce plausible-looking but conceptually debatable subclass relations or property assignments. In some cases, QS worsens after additional epochs, which may indicate overfitting to structural patterns specific to Q1 and reduced generalization to test ontologies. These observations suggest that while formal syntax and structural regularity are ef-

fectively learned through quality-guided continual pretraining, deeper semantic adequacy and robust generalization remain challenging objectives for future work.

4.2 Post-hoc Validation of the Quality Score

To better understand whether the proposed quality ranking captures properties genuinely beneficial for continual pretraining, we perform a lightweight post-hoc analysis relating ontology-level metric values to downstream ontology-generation performance. For each ontology, we compute Pearson correlations between QS, NTR, PD, and SC and a proxy of training utility derived from the average ontology-generation error of the best-performing epoch per architecture.

Metric	LLaMA	Gemma
QS	**-0.69** [-0.71, -0.66]	**-0.78** [-0.80, -0.76]
NTR	-0.59 [-0.62, -0.56]	-0.68 [-0.70, -0.65]
PD	-0.52 [-0.55, -0.48]	-0.64 [-0.67, -0.61]
SC	-0.37 [-0.41, -0.33]	-0.35 [-0.39, -0.31]

Table 5: Pearson correlations between ontology-level quality metrics and downstream ontology-generation performance (**mean error per triple**), reported as r [95% CI] on the full dataset ($N = 1766$). Negative values indicate that higher metric values are associated with lower generation error. All correlations are significant ($p < 0.001$). **QS**: Quality Score. **NTR**: Average Non-Taxonomic Relations per Class. **PD**: Property Density. **SC**: Average Subclasses per Class.

As reported in Table 5, the aggregated score QS exhibits the strongest negative correlation with ontology-generation error in both LLaMA and Gemma, outperforming all individual metrics. Since lower error indicates better performance, these negative correlations imply that **higher-quality ontologies are systematically associated with better downstream generation outcomes**. Correlation magnitudes are also stronger for Gemma than for LLaMA, suggesting that the extent to which a model exploits structural quality signals may depend on architectural or optimization factors.

This analysis should not be interpreted as a definitive validation of the optimality of QS: a full comparison would require additional continual pretraining runs with alternative metric combinations or weighting schemes. Nevertheless, the consistency of the correlation patterns across both architectures provides empirical support for the central methodological assumption of this work, namely that a simple aggregated score based on taxonomic structure, non-taxonomic connectivity, and property density can serve as an effective proxy for ontology usefulness in continual pretraining.

4.3 General NLP Performance

To assess broader effects beyond ontology generation, we evaluated LLaMA on a diverse set of open NLP benchmarks in English and Spanish (see Table 4). Due to the high computational cost of training multiple architectures, we selected LLaMA—the model that demonstrated superior ontology-generation performance—for this analysis.

Table 4 presents benchmark results for LLaMA across a diverse set of tasks in English and Spanish (Baucells et al., 2025). Overall, continual pretraining with ontological data shows **broad retention** of general-purpose NLP abilities, with several tasks even exhibiting improvements over the base model.

The benefits are most pronounced in tasks requiring **structured reasoning, semantic consistency, or inference**. For example, LLaMA improves on `xstorycloze_en/es`, `wnli/en-es`, `cola`, and `arc_challenge/easy`, all of which demand comprehension of concept relations, logical consistency, or commonsense reasoning. These gains are consistent with the model's exposure to highly structured symbolic ontologies during continual pretraining, suggesting that learning ontology regularities can transfer to reasoning-heavy NLP tasks.

In contrast, tasks involving **open-ended generation**, such as `xlsum_es` (summarization) and `phrases_es-va` (translation), show mild performance degradation. This indicates a modest **specialization effect**: while the model becomes more sensitive to structured, rule-governed generation, its capacity for unconstrained natural-language generation is slightly reduced.

An additional observation is that the subset yielding the best ontology-generation performance (Q1) does not always produce the highest scores across all general NLP tasks. Different quartile-based subsets (Q1, Q1-2, Q1-4) excel on different benchmarks, rein-

forcing that **ontology-specific adaptation and general NLP performance are related but distinct objectives**. High-quality semantic data is particularly effective for learning ontology-writing regularities, whereas benefits for broad NLP remain selective and task-dependent.

Overall, Table 4 demonstrates that **semantic continual pretraining can selectively enhance reasoning and structure-sensitive tasks** without causing catastrophic forgetting, confirming the practical value of ontology-guided adaptation even beyond ontology-generation objectives.

5 *Conclusions*

This work presents a quality-driven continual pretraining methodology for adapting compact open-source LLMs to ontology-to-ontology generation, making three main contributions:

- **Data-centric continual pretraining:** Ranking and segmenting ontological corpora to prioritize high-quality data over sheer volume for efficient LLM adaptation.
- **Controlled ontology-to-ontology evaluation:** Expert annotations and a fine-grained error taxonomy enable systematic benchmarking of model outputs.
- **Empirical analysis linking quality to performance:** Aggregated quality scores correlate strongly with generation accuracy and efficiency; using ∼31% of the data achieves the best results, reducing training time by up to 67% for one epoch (see Fig. 2).

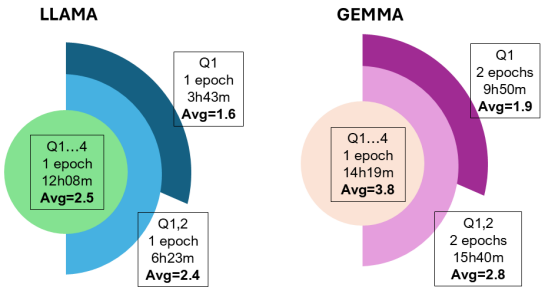

Figure 2: Training time to reach the best epoch per dataset segment and model. Full circle: all data (Q1–4); inner arcs: subsets (Q1–2, Q1). Time and **Avg error** shown; bold Avg indicates best model per architecture.

Experimental results show that training on the curated Q1 subset consistently outperforms larger, noisier corpora across both LLaMA and Gemma models, particularly in reducing repetition and syntactic instability. Evaluation on general NLP tasks indicates that semantic continual pretraining does not lead to severe catastrophic forgetting, with modest improvements observed in reasoning-oriented benchmarks and only slight degradation in summarization and translation tasks.

Overall, these findings emphasize that prioritizing dataset quality enhances ontology-generation performance, improves computational efficiency, and supports robust LLM adaptation in resource-constrained scenarios. This work highlights quality-driven corpus selection as a practical strategy for developing more efficient and sustainable pipelines for semantic and structured generation tasks.

Limitations

This study has some limitations. Experiments are restricted to 1B-parameter models, so whether quality-driven effects persist at larger scales or across different architectural families remains an open question. Although the ontology-generation evaluation is carefully designed and manually annotated, its moderate size limits coverage of the full diversity of possible ontology-continuation behaviors. In addition, the evaluation relies on an expert annotator, which provides a solid manual assessment of model outputs, although peer review or multiple annotators would further strengthen its statistical robustness.

The work also focuses exclusively on TTL-formatted ontologies and does not examine alternative semantic serializations. Moreover, models are adapted only through continual pretraining and are not subsequently instruction-tuned for ontology generation, which may limit alignment with task-specific generation goals. Finally, although the proposed QS score is empirically supported by post-hoc correlations, the current experimental design does not include a token-matched random baseline, which would help isolate the effect of ontology quality from data size and diversity, nor a full ablation over alternative metric combinations or weighting schemes.

Acknowledgments

This work was supported by the Spanish Ministry of Science, Innovation and Universities (MICIU) and the State Research Agency (AEI/10.13039/501100011033) through Grants HEART-NLP (PID2024-156263OB-C22) and PID2024-160791OB-I00, co-funded by ERDF/EU; Grant AIA2025-163322-C63, co-funded by ESF+; and Grants ACIE25-08, UAUSTI25-12, and ATI2025, funded by the Vice-Rectorate for Research of the University of Alicante.

References

[Babaei Giglou, D'Souza, y Auer2023] Babaei Giglou, H., J. D'Souza, y S. Auer. 2023. Llms4ol: Large language models for ontology learning. En *The Semantic Web – ISWC 2023*, páginas 408–427, Cham. Springer Nature Switzerland.

[Babaei Giglou et al.2025] Babaei Giglou, H., J. D'Souza, F. Engel, y S. Auer. 2025. Llms4om: Matching ontologies with large language models. En *The Semantic Web: ESWC 2024 Satellite Events*, páginas 25–35, Cham. Springer Nature Switzerland.

[Baucells et al.2025] Baucells, I., J. Aula-Blasco, I. de Dios-Flores, S. Paniagua Suárez, N. Perez, A. Salles, S. Sotelo Docio, J. Falcão, J. J. Saiz, R. Sepulveda Torres, J. Barnes, P. Gamallo, A. Gonzalez-Agirre, G. Rigau, y M. Villegas. 2025. IberoBench: A benchmark for LLM evaluation in Iberian languages. En O. Rambow L. Wanner M. Apidianaki H. Al-Khalifa B. D. Eugenio, y S. Schockaert, editores, *Proceedings of the 31st International Conference on Computational Linguistics*, páginas 10491–10519, Abu Dhabi, UAE, Enero. Association for Computational Linguistics.

[Belz y Reiter2006] Belz, A. y E. Reiter. 2006. Comparing automatic and human evaluation of NLG systems. En D. McCarthy y S. Wintner, editores, *11th Conference of the European Chapter of the Association for Computational Linguistics*, páginas 313–320, Trento, Italy, Abril. Association for Computational Linguistics.

[Chen et al.2019] Chen, H., G. Cao, J. Chen, y J. Ding. 2019. A practical framework for evaluating the quality of knowledge graph. En *China conference on knowledge graph and semantic computing*, páginas 111–122. Springer.

[Chen et al.2023] Chen, Z., Y. Liu, L. Chen, S. Zhu, M. Wu, y K. Yu. 2023. Opal: ontology-aware pretrained language model for end-to-end task-oriented dialogue. *Transactions of the Association for Computational Linguistics*, 11:68–84.

[Conneau et al.2020] Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, y V. Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. En *Proceedings of the 58th annual meeting of the association for computational linguistics*, páginas 8440–8451.

[Du et al.2024] Du, R., H. An, K. Wang, y W. Liu. 2024. A short review for ontology learning: Stride to large language models trend.

[Duque-Ramos et al.2011] Duque-Ramos, A., J. T. Fernández-Breis, R. Stevens, y N. Aussenac-Gilles. 2011. Oquare: A square-based approach for evaluating the quality of ontologies. *Journal of research and practice in information technology*, 43(2):159–176.

[Fathallah et al.2025] Fathallah, N., A. Das, S. D. Giorgis, A. Poltronieri, P. Haase, y L. Kovriguina. 2025. Neon-gpt: A large language model-powered pipeline for ontology learning. En *The Semantic Web: ESWC 2024 Satellite Events*, páginas 36–50, Cham. Springer Nature Switzerland.

[Frey et al.2020] Frey, J., D. Streitmatter, F. Götz, S. Hellmann, y N. Arndt. 2020. Dbpedia archivo: A web-scale interface for ontology archiving under consumer-oriented aspects. En *Semantic Systems. In the Era of Knowledge Graphs*, páginas 19–35, Cham. Springer International Publishing.

[Giglou et al.2025] Giglou, H. B., J. D'Souza, N. Mihindukulasooriya, y S. Auer. 2025. Llms4ol 2025 overview: The 2nd large language models for ontology learning challenge. *Open Conference Proceedings*.

[Gonzalez-Agirre et al.2025] Gonzalez-Agirre, A., M. Pàmies, J. Llop, I. Baucells, S. Da Dalt, D. Tamayo, J. J. Saiz,

F. Espuña, J. Prats, J. Aula-Blasco, y others. 2025. Salamandra technical report. *arXiv preprint arXiv:2502.08489*.

[Gururangan et al.2020] Gururangan, S., A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, y N. A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. En *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, páginas 8342–8360, Online, Julio. Association for Computational Linguistics.

[Gutierrez, Tomas, y Moreno2019] Gutierrez, Y., D. Tomas, y I. Moreno. 2019. Developing an ontology schema for enriching and linking digital media assets. *Future Generation Computer Systems*, 101:381–397.

[Huang et al.2025] Huang, W., G. Zhou, M. Lapata, P. Vougiouklis, S. Montella, y J. Z. Pan. 2025. Prompting large language models with knowledge graphs for question answering involving long-tail facts. *Knowledge-Based Systems*, 324:113648.

[Ibrahim et al.2024] Ibrahim, N., S. Aboulela, A. Ibrahim, y R. Kashef. 2024. A survey on augmenting knowledge graphs (kgs) with large language models (llms): models, evaluation metrics, benchmarks, and challenges. *Discover Artificial Intelligence*, 4(1):76.

[Jones y Galliers1995] Jones, K. S. y J. R. Galliers. 1995. *Evaluating Natural Language Processing Systems: An Analysis and Review*. Springer.

[Kudugunta et al.2023] Kudugunta, S., I. Caswell, B. Zhang, X. Garcia, D. Xin, A. Kusupati, R. Stella, A. Bapna, y O. Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. *Advances in Neural Information Processing Systems*, 36:67284–67296.

[Lantow2016] Lantow, B. 2016. Ontometrics: Putting metrics into use for ontology evaluation. En *KEOD*, páginas 186–191.

[Lippolis et al.2025] Lippolis, A. S., M. J. Saeedizade, R. Keskisärkkä, S. Zuppiroli, M. Ceriani, A. Gangemi, E. Blomqvist, y A. G. Nuzzolese. 2025. Ontology generation using large language models.

[LLaMA2024] LLaMA. 2024. Model cards and prompt formats – llama 3.2. `https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_2/`. Accessed: 2025-03-04.

[Lourdusamy y John2018] Lourdusamy, R. y A. John. 2018. A review on metrics for ontology evaluation. En *2018 2nd International conference on inventive systems and control (ICISC)*, páginas 1415–1421. IEEE.

[Lushnei et al.2026] Lushnei, S., D. Shumskyi, S. Shykula, E. Jiménez-Ruiz, y A. d. Garcez. 2026. Large language models as oracles for ontology alignment. En *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, páginas 2435–2449, Rabat, Morocco, Marzo. Association for Computational Linguistics.

[Monfardini, Salamon, y Barcellos2023] Monfardini, G. K. Q., J. S. Salamon, y M. P. Barcellos. 2023. Use of competency questions in ontology engineering: A survey. En *International Conference on Conceptual Modeling*, páginas 45–64. Springer.

[Palomar-Giner et al.2024] Palomar-Giner, J., J. J. Saiz, F. Espuña, M. Mina, S. Da Dalt, J. Llop, M. Ostendorff, P. O. Suarez, G. Rehm, A. Gonzalez-Agirre, y others. 2024. A curated catalog: Rethinking the extraction of pretraining corpora for mid-resourced languages. En *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, páginas 335–349.

[Pan et al.2026] Pan, C., Q. Yang, X. Zhang, S. Luo, Y. He, y J. Xie. 2026. The ontology of adverse events in 2025. *Scientific Data*.

[Pan et al.2024] Pan, S., L. Luo, Y. Wang, C. Chen, J. Wang, y X. Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599.

[Poveda-Villalón, Suárez-Figueroa, y Gómez-Pérez2012] Poveda-Villalón, M., M. C. Suárez-Figueroa, y A. Gómez-Pérez. 2012.

Validating ontologies with oops! En *International conference on knowledge engineering and knowledge management*, páginas 267–281. Springer.

[Raffel et al.2020] Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, y P. J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

[Saeedizade y Blomqvist2024] Saeedizade, M. J. y E. Blomqvist. 2024. Navigating ontology development with large language models. En *The Semantic Web*, páginas 143–161, Cham. Springer Nature Switzerland.

[Steinigen et al.2026] Steinigen, D., R. Teucher, T. H. Ruland, M. Rudat, N. Flores-Herr, P. Fischer, N. Milosevic, C. Schymura, y A. Ziletti. 2026. Fact finder - enhancing domain expertise of large language models by incorporating knowledge graphs. En *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, páginas 101–110, Rabat, Marocco, Marzo. Association for Computational Linguistics.

[Stevens et al.2019] Stevens, R., P. Lord, J. Malone, y N. Matentzoglu. 2019. Measuring expert performance at manually classifying domain entities under upper ontology classes. *Journal of Web Semantics*, 57:100469.

[Suárez, Sagot, y Romary2019] Suárez, P. J. O., B. Sagot, y L. Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. En *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

[Takahashi y Ishihara2025] Takahashi, H. y S. Ishihara. 2025. Quantifying memorization in continual pre-training with Japanese general or industry-specific corpora. En *Proceedings of the First Workshop on Large Language Model Memorization (L2M2)*, páginas 95–105, Vienna, Austria, Agosto. Association for Computational Linguistics.

[Team et al.2025] Team, G., A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, y M. Rivière. 2025. Gemma 3 technical report.

[Tello2002] Tello, A. J. L. 2002. *Métrica de idoneidad de ontologías*. Ph.D. tesis, Universidad de Extremadura.

[Toro et al.2024] Toro, S., A. V. Anagnostopoulos, S. M. Bello, K. Blumberg, R. Cameron, L. Carmody, A. D. Diehl, D. M. Dooley, W. D. Duncan, P. Fey, P. Gaudet, N. L. Harris, M. P. Joachimiak, L. Kiani, T. Lubiana, M. C. Muñoz-Torres, S. O'Neil, D. Osumi-Sutherland, A. Puig-Barbe, J. T. Reese, L. Reiser, S. M. Robb, T. Ruemping, J. Seager, E. Sid, R. Stefancsik, M. Weber, V. Wood, M. A. Haendel, y C. J. Mungall. 2024. Dynamic retrieval augmented generation of ontologies using artificial intelligence (dragon-ai). *Journal of Biomedical Semantics*, 15(1):19, Oct.

[van der Lee et al.2019] van der Lee, C., A. Gatt, E. van Miltenburg, S. Wubben, y E. Krahmer. 2019. Best practices for the human evaluation of automatically generated text. En *Proceedings of the 12th International Conference on Natural Language Generation*, páginas 355–368, Tokyo, Japan, Octubre–Noviembre. Association for Computational Linguistics.

[Vieira da Silva et al.2024] Vieira da Silva, L. M., A. Kocher, F. Gehlhoff, y A. Fay. 2024. On the use of large language models to generate capability ontologies. En *2024 IEEE 29th International Conference on Emerging Technologies and Factory Automation (ETFA)*, páginas 1–8.

[Xie, Aggarwal, y Ahmad2024] Xie, Y., K. Aggarwal, y A. Ahmad. 2024. Efficient continual pre-training for building domain specific large language models. En *Findings of the Association for Computational Linguistics: ACL 2024*, páginas 10184–10201, Bangkok, Thailand, Agosto. Association for Computational Linguistics.

[Xu et al.2022] Xu, J., X. Liu, J. Yan, D. Cai, H. Li, y J. Li. 2022. Learning to break the loop: Analyzing and mitigating repetitions for neural text generation. *Ad-*

vances in Neural Information Processing Systems, 35:3082–3095.

[Xue et al.2021] Xue, L., N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, y C. Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. En *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies*, páginas 483–498.

[Yang et al.2025] Yang, W., L. Some, M. Bain, y B. Kang. 2025. A comprehensive survey on integrating large language models with knowledge-based methods. *Knowledge-Based Systems*, 318:113503.

[Yılmaz, Tanyer, y Dikmen2025] Yılmaz, D., A. M. Tanyer, y I. Dikmen. 2025. Developing an ontology-based tool for relating risks to the energy performance gap in buildings. *Engineering, Construction and Architectural Management*.

[Zhang et al.2024] Zhang, B., V. A. Carriero, K. Schreiberhuber, S. Tsaneva, L. S. González, J. Kim, y J. de Berardinis. 2024. Ontochat: a framework for conversational ontology engineering using language models. En *European semantic web conference*, páginas 102–121. Springer.

[Zhao, Vetter, y Aryan2024] Zhao, Y., N. Vetter, y K. Aryan. 2024. Using large language models for ontoclean-based ontology refinement.