Do LLMs Know What Luxembourgish Borrows? Probing Lexical Neology in Low-Resource Multilingual Models

Nina Hosseini-Kivanani[*1,2]
[1] University of Luxembourg, Luxembourg
[2] Radio Télévision Luxembourg (RTL), Luxembourg
nina.hosseinikivanani@ext.uni.lu

Abstract

Large language models (LLMs) are increasingly used for writing assistance in small contact languages, yet it is unclear whether they respect community norms around lexical borrowing and neology. We introduce LexNeo-Bench, a 3,050-instance token-level benchmark derived from LuxBorrow, a large-scale Luxembourgish news corpus, where target tokens are labelled as native or as French, German, or English borrowings. Using this benchmark, we probe three multilingual LLMs across 34 prompt settings on two tasks: borrowing type classification and a binary lexical-innovation proxy (borrowing versus native). Without external context, models perform only slightly above chance on borrowing classification, so we construct a linguistic knowledge graph that encodes donor language, morphological patterns, and lexical analogues, and inject instance-specific subgraphs into the prompt. Knowledge-graph prompts raise borrowing classification accuracy from 25 – 35% up to 71 – 81% and largely close the gap between small and large models, while leaving neology detection difficult and sensitive to few-shot design. Our results show that lexicon-aware prompting is highly beneficial for robust borrowing judgments in low-resource contact languages and that lexical resources can serve as structured context for LLM evaluation. This study was carried out within the ENEOLI COST Action and examines borrowing as a form of lexical innovation in multilingual Luxembourgish data.

Keywords: Luxembourgish, lexical borrowing, neology, large language models, knowledge graphs

1. Introduction

Neology, the creation and diffusion of new lexical items, has long been central to lexicography, corpus linguistics, and sociolinguistics. With the emergence of large language models (LLMs), neology enters a new phase. LLMs are trained on massive multilingual corpora, absorb existing neologisms, and can themselves generate novel forms, blends, and hybrid structures in response to prompts. This raises questions not only about how LLMs detect and represent lexical innovation, but also about how their behavior interacts with existing norms and resources in individual language communities (Wolfer and Klosa-Kückelhaus, 2023; Zheng et al., 2024).

For smaller languages such as Luxembourgish, lexical innovation is tightly intertwined with contact phenomena. Luxembourgish exists in a dense contact zone with German, French, and English. Much of its modern lexical growth is realized through borrowing and adaptation from these donor languages rather than through entirely endogenous coinages (Adda-Decker et al., 2008). In written media, especially professionally edited news, many emergent forms are morphologically or orthographically integrated into Luxembourgish, while others remain closer to code-switching (Lavergne et al., 2014). For downstream Natural Language Processing (NLP) tools and LLM-powered applications, it matters whether these items are recognized as legitimate Luxembourgish words or treated as errors, foreign insertions, or targets for normalization back to French or German.

Previous work on Luxembourgish borrowing introduced LuxBorrow (Hosseini-Kivanani and Philippy, 2026), a large-scale corpus of Radio Télévision Luxembourg (RTL) news (1999–2025) annotated with sentence-level language identification and token-level labels for native items, borrowings from French, German, and English, and code-switching. That study focused on contact linguistic patterns and diachrony, showing that Luxembourgish remains the matrix language in news, while lexical borrowing and code mixing are pervasive but low-intensity, with a rich inventory of morphological and orthographic adaptation patterns. However, LuxBorrow did not address how contemporary LLMs treat these adapted forms, nor whether they recognize them as part of the Luxembourgish lexicon.

In this paper, we treat morphologically and orthographically adapted borrowings in Luxembourgish news as a key locus of lexical innovation and use LuxBorrow as ground truth to evaluate neology awareness in multilingual LLMs. We construct a token-level classification benchmark that pairs Luxembourgish sentences from RTL.lu with highlighted target tokens and gold labels indicating whether each token is native or a borrowing, and if so, from which donor language. On top of this benchmark, we define two tasks: a borrowing classification task in which models choose from

Proceedings of NEOLOGY AND LARGE LANGUAGE MODELS, pages 27–38
16 MAY 2026.

four labels (NATIVE, FR_LOAN, DE_LOAN, and EN_LOAN as a diagnostic distractor) but are evaluated on three gold classes, and a binary neology decision task.

Our study is organized around three research questions.

- RQ1. To what extent do off-the-shelf multilingual LLMs correctly classify native Luxembourgish words and distinguish French- vs German-origin adapted borrowings in RTL news?
- RQ2. Do LLMs systematically bias their judgments toward dominant donor languages, especially French and German, and how often do they incorrectly project English-origin hypotheses via the EN_LOAN distractor label?
- RQ3. How does providing explicit lexicon-based context, for example, a loanword registry, affect LLM performance and their treatment of Luxembourgish lexical innovation?

To answer these questions, we evaluate three strong multilingual LLMs in frozen, prompt-only mode. We compare zero-shot prompting, few-shot prompting with manually chosen examples of Luxembourgish borrowings, and two knowledge-based prompting conditions: `KG_flat`, which provides a global list of borrowing patterns, and `KG_graph`, which injects an instance-specific lexicon context derived from the LuxBorrow loanword registry.

Our contributions are threefold. First, we introduce LexNeo-Bench, a token-level benchmark for borrowing classification in Luxembourgish, derived from LuxBorrow, with public scripts for extraction, prompting, and evaluation. Second, we add a binary lexical-innovation proxy task that collapses borrowings versus native items to probe neology awareness, and show that it remains challenging even for strong multilingual LLMs. Third, we show that lightweight lexicon-based context via a linguistic knowledge graph can substantially improve borrowing judgments in a low-resource contact language, which suggests concrete avenues for integrating community-curated lexical resources into LLM prompting for neology-sensitive applications. Within the ENEOLI COST Action, this study contributes to WG2 by treating borrowing in Luxembourgish as a corpus-based case of lexical innovation and by evaluating how multilingual LLMs analyze such forms in a low-resource contact setting.

2. Related Work

2.1. Borrowing, code-switching, and neology

Contact linguistics distinguishes lexical borrowing, items integrated into the recipient language's lexicon and grammar, from code-switching, that is, spontaneous alternation between languages within discourse. Classic accounts emphasize that entrenched borrowings are morphologically and phonologically integrated, frequent, and often listed in dictionaries, while code-switches retain donor language structure and remain more speaker-specific. This view underlies the "Simple View" of borrowing, which operationalizes the difference in terms of listedness in the mental lexicon and community entrenchment (Treffers-Daller, 2025; Chesley and Baayen, 2010).

In multilingual European contexts, written media often show a stable matrix language with pervasive but shallow insertions from donor languages. Borrowings can be introduced via institutional domains such as politics, finance, and administration, before diffusing into more general registers. Over time, morphologically adapted forms may compete with native synonyms or with less integrated loan variants. This dynamic is particularly visible in Luxembourgish, where French and German both supply a rich inventory of technical and everyday lexical items, and where orthographic and morphological adaptation blur the surface boundary between native and borrowed forms (Anastasiou, 2022; Lavergne et al., 2014; Adda-Decker et al., 2008).

Lexicographic and corpus-based studies of neology therefore give prominence to borrowed and adapted items when tracking lexical innovation, especially in small languages that rely heavily on lexical importation from regional lingua francas (Wolfer and Klosa-Kückelhaus, 2023).

2.2. Computational borrowing and neology detection

In NLP, early work on multilingual text mixing emphasized document- or utterance-level indices, such as code-mixing indices and entropy-based measures, which treat all foreign tokens uniformly. More recent studies move to explicit borrowing detection and distinguish unassimilated foreign tokens, code-switches, and integrated loanwords. This line of work has introduced borrowing-annotated corpora, for example anglicism detection in Spanish newswire (Alvarez-Mellado, 2020, 2021), and shared tasks with sequence tagging baselines (Mellado et al., 2021; Álvarez-Mellado et al., 2025). Methods range from conditional random fields and BiLSTM-CRFs to

transformer taggers that incorporate lexical and orthographic features (Alvarez-Mellado, 2020; Álvarez Mellado, 2020), alongside resource-lean approaches to code-switching identification that rely mainly on word lists and monolingual corpora (Kevers, 2022).

Beyond borrowing per se, neology detection has traditionally relied on dictionary versus corpus comparisons combined with temporal information, for example, locating forms that appear in recent corpora but are absent from older lexica. With the advent of LLMs, recent work has begun to integrate these models into neologism detection pipelines, for example using them as filters or validators for candidate neologisms, and to provide lemmata and definitions for emergent forms (Tomaszewska et al., 2025; Hosseini-Kivanani, 2025). Other studies highlight how LLMs can also generate non-attested "LLM neologisms" due to tokenization and encoding artifacts (Iwamoto and Kanayama, 2024). This opens a new evaluation axis: not just whether LLMs can help detect neology, but whether their intrinsic lexical knowledge and biases align with community norms and lexicographic resources (Tomaszewska et al., 2025; Hosseini-Kivanani, 2025; Iwamoto and Kanayama, 2024).

2.3. LLMs, low resource languages, and lexical inequality

Work on LLMs in low-resource languages highlights skewed coverage and performance gaps, where models trained mainly on high-resource languages underrepresent or mis-analyze items from smaller languages and can "normalize" adapted borrowings back to donor forms. This has consequences for spell-checkers, assistive writing tools, and generation systems that interact with speakers of contact languages. Empirical studies of Luxembourgish resources and their multilingual context document sparse written production and heavy code-mixing and adaptation pressure, which exacerbate LLM coverage problems (Plum et al., 2024; Lavergne et al., 2014; Adda-Decker et al., 2008).

Lexicon-aware prompting and retrieval/gazetteer augmentation show that injecting compact community resources into LLM workflows yields large gains on complex Named Entity Recognition (NER) and entity-centric tasks (Tan et al., 2023; Chen et al., 2022). This motivates using curated loanword registries or structured knowledge-graph hints to probe LLM judgments about borrowed and adapted forms. Against this background, we use a borrowing-annotated Luxembourgish news corpus as a neology resource to build LexNeo-Bench, a benchmark that probes LLM lexical decisions in a dense contact setting.

3. Experiments

Figure 1 summarizes the overall evaluation pipeline, from LuxBorrow-derived benchmark construction and LKG retrieval to prompt assembly and multilingual LLM evaluation.

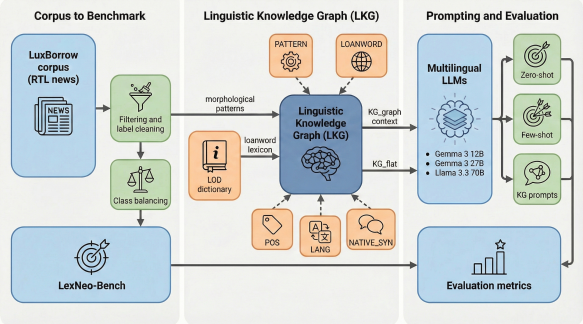

Figure 1: LexNeo-Bench pipeline.

3.1. Benchmark construction

We construct LexNeo-Bench, a token-level evaluation benchmark derived from the LuxBorrow corpus of professionally edited Luxembourgish news. LuxBorrow provides sentence-level language identification over RTL articles together with token-level borrowing labels generated by a morphological pattern pipeline. Each benchmark instance consists of a Luxembourgish sentence, a highlighted target token, its gold borrowing label, compact morphological evidence, and article metadata such as section and timestamp.

The label space follows the LuxBorrow taxonomy and distinguishes native Luxembourgish items (**NATIVE**) from French-, German-, and English-origin borrowings (**FR_LOAN**, **DE_LOAN**, **EN_LOAN**). Tokens tagged as **CODE_SWITCH** or as named entities are excluded from the main task, since the focus is on entrenched lexical items rather than span-level alternation. To avoid extremely sparse classes, we discard labels with fewer than 50 instances in the source corpus. EN_LOAN appears only 24 times and is therefore removed from the evaluation set, which yields a three-way task over NATIVE, FR_LOAN, and DE_LOAN, even though the conceptual four-way taxonomy is kept in the prompts.

The original LuxBorrow corpus comprises $259\,305$ RTL news articles and 43.7 million tokens. We first remove punctuation, a curated list of Luxembourgish function words, and tokens with low-confidence automatic labels. From the remaining pool, we draw 1000 instances per active class,

which results in a balanced benchmark of 3 000 examples, and we add a small diagnostic stratum of 50 CODE_SWITCH tokens for error analysis. The final benchmark therefore contains 3050 instances: NATIVE $= 1000$, FR_LOAN $= 1000$, DE_LOAN $= 1000$, and CODE_SWITCH $= 50$.

Each instance inherits the publication date of its source article (1999–2025), providing a diachronic signal. As a proxy for entrenchment, we contrast tokens from articles published before 2015 with those after 2015. This split is not first-attestation dating, but leverages a 25-year professionally edited news record. The borrowing labels serve as the primary gold standard for both tasks. For the auxiliary neology decision task, we derive a binary label at prompt construction time by mapping tokens annotated as `FR_LOAN`, `DE_LOAN`, or `EN_LOAN` to `YES` (lexical innovation) and `NATIVE` tokens to `NO`. The recent versus established flag is used only for temporal robustness analysis.

For prompt types that embed full lexicon entries, we restrict ourselves to a reduced subset of 674 benchmark items for which dictionary definitions, etymology, and related lexical information are available in a consistent format from the Lëtzebuerger Online Dictionnaire (LOD) (Zenter fir d'Lëtzebuerger Sprooch, 2025). This avoids noisy or incomplete context in lexicon-assisted conditions.

3.2. Linguistic Knowledge Graph

To provide models with structured linguistic context, we construct a **Linguistic Knowledge Graph (LKG)** that integrates three LuxBorrow-related resources: a compiled index of productive morphological patterns, a loanword lexicon extracted from LOD, and a hand-curated table of loanword–native synonym pairs.

Nodes are typed as morphological patterns (PATTERN), donor languages (LANG), loanword entries (LOANWORD), native Luxembourgish synonyms (NATIVE_SYN), and part-of-speech tags (POS). Edges encode linguistic relations such as pattern membership (*follows_pattern*), donor origin (*from_donor*), synonymy (*has_synonym*), competition between patterns with the same Luxembourgish affix but different donor languages (*contrastive*), and lexical category (*has_pos*).

For each benchmark instance, we perform multi-hop retrieval on this graph to construct a compact, token-specific explanation subgraph. Starting from the target token, we match compatible LuxBorrow patterns, query LOD to obtain donor metadata, and reconstruct an etymology-style chain of the form donor form $\rightarrow$ adaptation pattern $\rightarrow$ Luxembourgish form, collect linked native synonyms, and sample a small set of analogues that share a pattern, plus a few contrastive patterns with the same affix but different donor languages. The retrieved subgraph is then linearized into a structured natural-language block of at most 30 lines and prepended to the model prompt. This instance-specific `KG_graph` context replaces a much coarser `KG_flat` baseline in which the same global list of 19 patterns is appended to every example independently of the target token.

3.3. Prompt setups

All prompts share a common two-role template. The system message defines a Luxembourgish linguistics expert persona, and the user message concatenates optional knowledge-graph context, the task instruction, and the Luxembourgish sentence with the target token marked by `**`. The user message then introduces any external context, followed by a concise instruction to assign exactly one label to the highlighted token and the sentence in which it appears.

We evaluate a family of prompt strategies for both borrowing classification and neology detection. In all cases, the model receives a short English instruction, the Luxembourgish sentence, and the target token marked with `**`. For the **classification** task, the model must output exactly one label from the conceptual four-way set $\{$`NATIVE`, `FR_LOAN`, `DE_LOAN`, `EN_LOAN`$\}$. Although `EN_LOAN` does not appear in the evaluation data, we keep it as a possible answer to capture uncertainty toward English-origin candidates. For the neology task, the model must answer `YES` if the token should be treated as a lexical innovation in Luxembourgish, and `NO` otherwise.

The base prompt strategies include a plain `zero_shot` condition (system role plus task description, no additional context), a `few_shot` variant with five manually authored demonstrations, and a `minimal` variant that reduces the instruction to a single line and enforces label-only output. The five demonstrations do not overlap with the 3 050 benchmark instances. They consist of two prototypical `FR_LOAN` examples and one example each for `DE_LOAN`, `EN_LOAN`, and `NATIVE`. Each demonstration pairs a short Luxembourgish sentence with its gold label and a brief linguistically motivated justification.

An excerpt of the few-shot prompts is shown in Listing 1.

Listing 1: Excerpt of few-shot prompts for borrowing classification and neology decision.

```
System:
  You are a linguistic expert specializing in
     Luxembourgish
  (Lëtzebuergesch). Luxembourgish is a West
     Germanic language spoken
  in Luxembourg that regularly borrows and
     morphologically adapts
```

```
  words from French, German, and English.

User (classification task):
  Given the following Luxembourgish sentence and
     the highlighted
  word (marked with ** **), decide whether the
     highlighted word is:
    - NATIVE:  a native Luxembourgish word
    - FR_LOAN: a borrowing from French (
     morphologically adapted
                into Luxembourgish)
    - DE_LOAN: a borrowing from German (
     morphologically adapted
                into Luxembourgish)
    - EN_LOAN: a borrowing from English (
     morphologically adapted
                into Luxembourgish)
  Respond with ONLY the label on the first line and
      a one-sentence
   justification on the second line.

  Example:
    Sentence: D'**Pompjeeën** hunn de Brand
     schnell ënnert
              Kontroll bruecht.
    Assistant: FR_LOAN
    Justification: 'Pompjeeën' derives from French
     'pompier',
      adapted with the Luxembourgish plural suffix
     "-en" and
      spelling "ee" for /e:/.

User (neology task):
  Given the following Luxembourgish sentence and
     the highlighted
  word, decide whether this token should be
     treated as a lexical
  innovation in Luxembourgish.
  Answer YES or NO, followed by one sentence of
     explanation.

  Sentence: [Luxembourgish sentence containing **
     TOKEN**].
```

Knowledge-augmented prompts add morphological information derived from LuxBorrow and LOD. In the `KG-flat` conditions, the user message begins with a preamble *"According to the LOD, the following morphological adaptation patterns are productive in Luxembourgish:"*, followed by up to twenty globally fixed pattern entries. Each entry lists a pattern name, its type (morphological, orthographic, or lexical), the donor language, and up to three example pairs, for example *"éiere → er, type. morph, donor. FR, e.g. abordéieren ← aborder"*. This global pattern block is identical in `KG-flat` and is appended to every instance, which contrasts with `KG-graph`, where the context is an instance-specific LKG subgraph as described in Section 3.2.

To quantify the contribution of individual LKG components, we define six ablation variants that selectively remove lexicon attestation, etymology chains, synonym links, analogical examples, or contrastive patterns, as well as a `lex-only` condition that keeps only dictionary-style information without graph structure. Together, the eleven base strategies and six ablations define 17 prompt setups per task. Applied to both borrowing classification and neology detection, this yields 34 task-specific evaluation settings per model.

3.4. Models

All experiments are conducted with instruction-tuned, general-purpose LLMs accessed through an OpenAI-compatible endpoint (OpenRouter API). We deliberately treat the models as frozen black boxes and rely exclusively on prompting; no fine-tuning is performed.

We consider three model sizes: Gemma 3 12B (`google/gemma-3-12b-it`) as a small model, Gemma 3 27B (`google/gemma-3-27b-it`) as a medium model, and Llama 3.3 70B Instruct (`meta-llama/llama-3.3-70b-instruct`) as a large model. All runs use a temperature of 0.0 and a maximum output length of 1,024 tokens to enforce deterministic, label-complete responses. Combining three models with 34 prompt configurations yields 102 evaluation settings, each applied to the full LexNeo-Bench of 3 050 instances.

3.5. Evaluation protocol

Model outputs often contain explanations or formatting artifacts, so we post-process responses to recover a single canonical label per instance. We strip explicit reasoning blocks (for example between `<think>` and `</think>`), then examine the first and last non-empty lines and map them to one of the allowed labels using a small normalization dictionary (for example, `FRENCH`, `FR`, or `FR_loanword` all map to `FR_LOAN`, while `LUXEMBOURGISH` or `LB` map to `NATIVE`). Outputs that cannot be unambiguously resolved are marked as `PARSE_ERROR` and omitted from metric computation; we report their frequency separately.

For the borrowing classification task, we report accuracy, balanced accuracy, macro- and weighted-F1 over the active classes, as well as per-class precision, recall, F1, and confusion matrices. In addition, we analyze two derived sub-tasks: a binary native versus borrowed decision (collapsing `FR_LOAN` and `DE_LOAN`) and donor-only discrimination between `FR_LOAN` and `DE_LOAN`. For the neology task, we treat `YES` as the positive label and report accuracy, precision, recall, and F1, with additional breakdowns by donor language. The gold label is derived directly from the primary LuxBorrow borrowing annotation: tokens annotated as `FR_LOAN`, `DE_LOAN`, or `EN_LOAN` are mapped to `YES` (lexical innovation), and `NATIVE` tokens to `NO` (see Section 3.1). All metrics are computed on the same fixed test set.

Temporal robustness is assessed by comparing accuracies on established versus recent items and reporting the absolute gap. Finally, the evaluation pipeline supports resumable execution. Prediction files are incrementally extended when experiments are restarted, which makes large grids of runs robust to interruptions without recomputation.

4. Results

4.1. RQ1. Borrowing classification performance

Table 1 summarizes three-way borrowing classification accuracy and macro F1 across models and prompt strategies. Without a structured linguistic context, performance remains modest. In the zero-shot baseline, accuracy ranges from 24.5% for Gemma 3 12B to 34.7% for Llama 3.3 70B, and more elaborate non-KG prompts, such as Few-shot, remain below 42% across all models.

Since models choose from four output labels, a random baseline yields 25% accuracy; zero-shot performance ranges from 24.5% to 34.7%, indicating that parametric knowledge alone barely exceeds chance.

Introducing a structured linguistic context via the KG-graph condition changes this picture sharply. With KG-graph, accuracy rises to 81.0% for Gemma 3 12B, 71.4% for Gemma 3 27B, and 71.3% for Llama 3.3 70B, and macro F1 exceeds 0.55 for all models, peaking at 0.634 for Gemma 3 12B. The simpler KG-flat variant, which exposes only a global list of morphological patterns, does not close this gap and behaves similarly to non-KG baselines. The improvement, therefore, stems from instance-specific retrieval rather than merely reminding the model that borrowing patterns exist. Taken together, these results answer RQ1 by showing that structured, token-level linguistic context is necessary to achieve robust borrowing classification in Luxembourgish.

Table 1: Acc. and macro F1 (in %), and KG gain Δ_{KG} (percentage points), defined as the accuracy difference between KG-graph and zero-shot.

Model	Prompt	Acc.(%)	Macro F1	Δ_{KG}
Gemma 3 12B	Zero-shot	24.5	22.3	
	Few-shot	38.3	30.3	
	KG-flat	30.3	26.2	
	KG-graph	81.0	63.4	+56.5
Gemma 3 27B	Zero-shot	31.4	19.7	
	Few-shot	38.0	29.8	
	KG-flat	33.7	22.6	
	KG-graph	71.4	55.9	+40.1
Llama 3.3 70B	Zero-shot	34.7	27.9	
	Few-shot	38.0	29.0	
	KG-flat	36.3	27.9	
	KG-graph	71.3	55.7	+36.6

4.2. RQ2. Per class performance and donor bias

To understand where the gains from KG-graph conditioning arise, Figure 2(A) reports per class F1 under the KG-graph prompt. All three models achieve strong F1 scores for French and German borrowings. Gemma 3 12B reaches 0.920 for FR_LOAN (French borrowing) and 0.840 for DE_LOAN (German borrowing); Gemma 3 27B reaches 0.880 and 0.750 respectively, and Llama 3.3 70B scores 0.921 and 0.791. Performance on NATIVE items is more variable. Gemma 3 12B maintains a solid 0.777 F1, whereas Llama 3.3 70B drops to 0.515, suggesting that the larger model overfits to donor cues and sometimes over-predicts borrowing for genuinely native words.

Confusion patterns show a marked donor asymmetry. In Figure 2(b), the dominant error is French-origin items misclassified as German-origin (FR_LOAN→DE_LOAN: 18,142 cases across all model and prompt combinations), which is 4.6× more frequent than the reverse direction (DE_LOAN→FR_LOAN: 3,946). Native Luxembourgish items are also misattributed to German borrowings (20,662) substantially more often than to French borrowings (5,944), indicating an overall tendency to overpredict DE_LOAN. This pattern is consistent with potential lexical/orthographic overlap between French- and German-origin forms in Luxembourgish, although other factors (e.g., class priors or KG coverage) may also contribute. Overall, these results support RQ2: while KG-graph improves borrowing recognition, donor identification remains skewed toward German across model and prompt settings.

EN_LOAN as a distractor label. Although EN_LOAN is absent from the evaluation set (only 24 source instances, below the 50-instance threshold), we retain it as a valid output label to probe whether models project English-origin hypotheses onto tokens that are in fact native or borrowed from French or German. Table 2 reports how often each model predicts EN_LOAN and which true class absorbs those false positives.

Table 2: EN_LOAN false-positive analysis. **EN_{pred}** is the total number of EN_LOAN predictions; columns show the true-class breakdown of those predictions. **Rate** is the proportion of all valid predictions assigned to EN_LOAN.

Model	Prompt	EN_{pred}	→NAT	→FR	→DE	Rate
Gemma 12B	Zero-shot	1 001	392	196	387	32.8%
	Few-shot	238	103	29	97	7.8%
	KG-flat	627	254	80	275	20.6%
	KG-graph	128	107	5	12	4.2%
Gemma 27B	Zero-shot	312	130	47	129	10.2%
	Few-shot	212	96	23	89	7.0%
	KG-flat	299	125	31	136	9.8%
	KG-graph	167	128	6	28	5.5%
Llama 70B	Zero-shot	363	196	37	125	12.0%
	Few-shot	91	38	6	43	3.0%
	KG-flat	87	32	7	45	2.9%
	KG-graph	264	222	7	32	8.7%

Two patterns stand out. First, without structured context, models frequently over-predict EN_LOAN: Gemma 12B assigns it to nearly a third of all

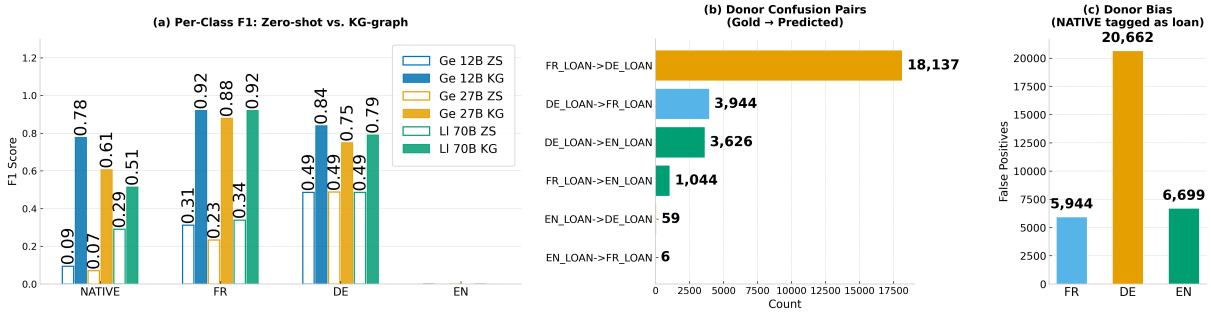

Figure 2: (a) Per-class F1 for zero-shot and KG-graph by model (GE12, GE27, & LL70 denote Gemma 3 12B, Gemma 3 27B, & Llama 3.3 70B, respectively.). (b) Top donor confusion pairs. (c) False-positive rates on NATIVE: proportion of NATIVE tokens predicted as loans.

tokens under zero-shot prompting. KG-graph prompting reduces the EN_LOAN rate by 78–87% for the Gemma models (from 32.8% to 4.2% for Gemma 12B, and from 10.2% to 5.5% for Gemma 27B), confirming that structured linguistic context suppresses spurious English-origin hypotheses. Second, across all models and prompt conditions, the majority of false EN_LOAN predictions fall on genuinely NATIVE tokens rather than on French or German borrowings. Under KG-graph, this concentration intensifies: 84% of Gemma 12B's and 77% of Gemma 27B's residual EN_LOAN predictions fall on NATIVE tokens. This suggests that when models lack donor-specific evidence, they default to an English-origin hypothesis for unfamiliar Luxembourgish words, a bias consistent with English's dominance in multilingual pretraining corpora.

Retaining EN_LOAN as a distractor label therefore serves a diagnostic purpose: it exposes this bias and provides a measurable signal of how effectively structured context can counteract it.

4.3. RQ3. Ablating KG components

Figure 3 shows the effect of removing individual components from the KG-graph prompt. The full KG-graph condition reaches 81.0%, 71.4%, and 71.3% accuracy for the three models (Gemma 3 12B, Gemma 3 27B, and Llama 3.3 70B). Removing etymological information (No Etymology) reduces accuracy to 78.9% for Gemma 3 12B, 58.4% for Gemma 3 27B, and 69.2% for Llama 3.3 70B. Dropping analogical examples (No Analogues) has a similarly strong impact, especially on the 27B model, where accuracy decreases by roughly 13 percentage points.

By contrast, removing synonym links or contrastive patterns changes performance only marginally, within ± 0.3 points of the full KG-graph condition. A Lexicon-only variant that keeps dictionary entries but discards graph structure clearly outperforms non-KG baselines, yet remains 6–19 points behind the full graph, which suggests that donor chains and pattern-sharing analogues carry most of the useful signal, while long definitions may introduce noise. In some settings, accuracy even improves slightly when the lexicon text is removed, but the graph structure is kept, reinforcing that relational structure is more valuable than raw definitional prose.

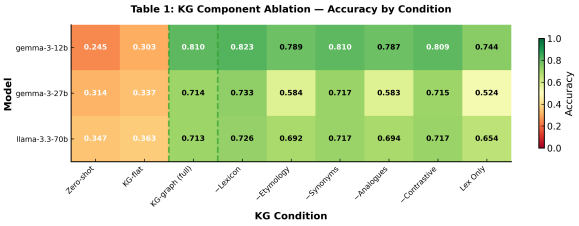

Figure 3: Impact of KG components on borrowing accuracy by model and KG-graph ablation condition.

4.4. Model scale and benefit from KG

Table 1 compares zero-shot and KG-graph accuracy by model size. Zero-shot accuracy grows modestly with scale, from 24.5% for Gemma 3 12B to 34.7% for Llama 3.3 70B, but under KG-graph the ranking inverts: Gemma 3 12B reaches 81.0%, while Gemma 3 27B and Llama 3.3 70B plateau at 71.4% and 71.3%, respectively. The KG gain Δ_{KG}, defined as the accuracy difference between KG-graph and zero-shot, decreases monotonically with model size: +56.5, +40.1, and +36.6 percentage points. A supplementary log-scale visualization of this trend is provided in Appendix 4.

The reversal is KG-specific. Gemma 12B does *not* generally outperform Gemma 27B. Under zero-shot (31.4% vs. 24.5%), few-shot (38.0% vs. 38.3%, essentially tied), and KG-flat (33.7% vs. 30.3%), Gemma 27B matches or exceeds

Gemma 12B. The reversal occurs exclusively under KG-graph (+9.5 pp in favor of 12B), ruling out a general quality advantage of the smaller model and localizing the effect to how each model utilizes instance-specific structured context.

Mechanism: NATIVE over-prediction by larger models. Under KG-graph, all models achieve high recall on FR_LOAN ($\geq$0.860) and DE_LOAN ($\geq$0.961), but NATIVE recall drops sharply with scale: 0.639 (12B), 0.447 (27B), 0.349 (70B). Precision on NATIVE remains above 0.94 for all models, so larger models predict NATIVE correctly when they do—but they predict it far too rarely, over-attributing borrowing status to native words. KG ablations confirm this asymmetry: removing etymology or analogues costs Gemma 27B $\sim$13 pp but Gemma 12B only $\sim$2 pp, showing that the larger model falls back on parametric borrowing priors when graph evidence is incomplete.

This pattern is consistent with findings on parametric–contextual knowledge conflicts (Longpre et al., 2021; Xie et al.): larger models develop stronger internal representations of French and German items during pre-training, which compete with KG-supplied evidence and lead to over-attribution of donor origins. The smaller model, lacking such entrenched priors, defers more faithfully to the structured context.

In this analysis, we use publication dates as a diachronic proxy to contrast more established items with more recent adaptations (pre-2015 vs. post-2015). The graph encodes structural origin information (donor language, morphological patterns, analogues) but not explicit recency cues such as frequency trajectories or first-attestation dates. Under this temporal split (see Supplementary), KG-graph is the most temporally robust condition, with recent vs. established gaps of only 0.7–2.8 pp.

Under the KG-graph condition, Gemma 3 12B achieves 81.4% accuracy for established items and 80.7% for recent ones; Gemma 3 27B achieves 73.1% and 70.3%; Llama 3.3 70B reaches 72.4% and 70.6%. These results indicate that the graph captures structural regularities that transfer to more recent lexical items, even if such items are under-represented or missing in the models' pre-training data. Among all prompt strategies, KG-graph is the least affected by recency, suggesting that structured linguistic context can partially compensate for gaps in parametric training data; a full breakdown by model and prompt is provided in the supplementary material.

4.5. Neology detection

The neology decision task, which collapses all borrowings into a single lexical-innovation class versus native items, behaves very differently from borrowing classification. Table 3 reports accuracy and $F1_{\text{neo}}$ for the "neologism" class by model and prompt strategy. Here, few-shot prompting is consistently the most effective strategy. Gemma 3 12B reaches 48.5% accuracy and $F1_{\text{neo}} = 0.509$, Gemma 3 27B reaches 49.2% and 0.524, and Llama 3.3 70B achieves 40.8% and 0.254. In contrast, the KG-graph condition substantially degrades performance. Accuracy falls to 34.2%, 30.3%, and 30.5% for the three models, and $F1_{\text{neo}}$ for Llama 3.3 70B drops close to zero.

This divergence is in line with how the linguistic knowledge graph is constructed. The graph encodes origin and structural information (donor language, morphological pattern, analogues, native synonyms), which are exactly the cues needed for borrowing classification, but largely orthogonal to *recency*. Deciding whether a word counts as a lexical innovation requires diachronic evidence, such as frequency trajectories, first attestation dates, or domain-specific usage shifts, none of which are currently exposed in the graph. As a result, the additional context encourages models to reason about *where* a word comes from rather than *when* it entered the language, which can mislead them in borderline cases.

Table 3: Neology decision performance by model and prompt. Accuracy and $F1_{\text{neo}}$ for the "neologism" class.

Model	Prompt	Acc. (%)	$F1_{\text{neo}}$
Gemma 3 12B	Zero-shot	41.2	0.308
	Few-shot	48.5	0.509
	KG-graph	34.2	0.042
Gemma 3 27B	Zero-shot	45.6	0.429
	Few-shot	49.2	0.524
	KG-graph	30.3	0.064
Llama 3.3 70B	Zero-shot	36.7	0.127
	Few-shot	40.8	0.254
	KG-graph	30.5	0.012

4.6. Binary native versus borrowed

Finally, we collapse the four-class label space into a binary decision and ask models to distinguish native Luxembourgish words from any type of borrowing. Under the KG-graph condition, all models reach high performance. Gemma 3 12B attains 85.5% accuracy and $F1 = 0.902$, Gemma 3 27B reaches 78.9% and 0.862, and Llama 3.3 70B reaches 75.5% and 0.845.

The contrast between the binary decision and the donor-specific four-way task suggests that the main residual difficulty lies in separating French from German borrowings, rather than in detecting

whether a token is lexically integrated at all. In other words, once the knowledge graph is available, knowing that a word is a borrowing is comparatively easy, while pinpointing the correct donor in a dense Luxembourgish, French, and German contact zone remains challenging. Detailed binary results for all prompt strategies are reported in the supplementary material.

5. Discussion

Our results show that off-the-shelf multilingual LLMs have limited awareness of how a small contact language integrates lexical borrowings, even when trained on large multilingual corpora. With four possible output labels, a random baseline yields 25% accuracy; zero-shot performance ranges from 24.5% to 34.7%, indicating that parametric knowledge alone barely exceeds chance. This observation aligns with work in contact linguistics and neology that emphasizes community entrenchment, dictionary listedness, and usage patterns over purely formal cues (Treffers-Daller, 2025; Chesley and Baayen, 2010; Wolfer and Klosa-Kückelhaus, 2023). The models do not spontaneously replicate the "Simple View" of borrowing as operationalized in lexicographic and corpus studies.

LexNeo-Bench complements earlier borrowing and anglicism corpora in Spanish and other languages (Alvarez-Mellado, 2020, 2021; Mellado et al., 2021; Álvarez-Mellado et al., 2025; Álvarez Mellado, 2020; Kevers, 2022) by exposing LLMs to a dense Luxembourgish, French, and German contact zone where orthographic and morphological integration is pervasive (Adda-Decker et al., 2008; Lavergne et al., 2014; Anastasiou, 2022). The strong gains from structured knowledge-graph prompting suggest that models can make fine-grained borrowing decisions once they are supplied with token-specific morphological patterns, donor labels, and analogical examples. This mirrors gains observed when injecting gazetteers and knowledge bases into NER and entity-centric tasks (Tan et al., 2023; Chen et al., 2022) and supports the view that community lexical resources remain crucial even in the LLM era (Tomaszewska et al., 2025; Hosseini-Kivanani, 2025).

At the same time, our neology decision results highlight that structural donor information alone does not solve diachronic questions. LLMs perform best with few-shot prompting that clarifies the task mapping (borrowings count as lexical innovations), while knowledge-graph prompts, which were designed for borrowing classification, can even harm performance. This gap reflects broader findings on LLM-based neology detection and "LLM neologisms" that arise from tokenization and encoding artifacts rather than organic community usage (Iwamoto and Kanayama, 2024; Zheng et al., 2024). For small languages with sparse written production and heavy code mixing (Plum et al., 2024; Adda-Decker et al., 2008), separating genuine innovations from long-standing borrowings remains challenging without explicit temporal signals or external diachronic corpora.

Our study has several limitations: First, LexNeo-Bench is derived from a single edited news source, so it under-represents informal registers and spoken discourse. Second, borrowing labels rely on an automatic pattern pipeline and dictionary signals, which may misclassify borderline items or miss emerging forms in under-documented domains. Third, we evaluate only three instruction-tuned models with frozen prompts, so conclusions about model scale and architecture should be treated as tentative. Finally, the benchmark focuses on token-level decisions and does not directly measure how LLMs handle borrowing in generation, for example, in spelling correction or style transfer. Addressing these limitations will require extending the benchmark to other genres, adding human validation for difficult cases, and coupling classification with controlled generation tasks.

6. Conclusion and Future Work

We introduced LexNeo-Bench, a token-level benchmark derived from a borrowing-annotated Luxembourgish news corpus to probe how multilingual LLMs treat morphologically adapted borrowings. Across three models and 34 prompt configurations, zero-shot parametric knowledge stays near chance, whereas instance-specific linguistic knowledge graphs raise borrowing classification accuracy to about 71–81% and substantially improve binary native versus borrowed decisions, with the largest gains for the smallest model. This shows that structured lexical context can partly compensate for sparse pretraining in low-resource contact languages, while neology decisions remain difficult and are best supported by few-shot prompting in our experiments because recency is not encoded in the current graph. Future work will add explicit diachronic signals, extend LexNeo-Bench beyond edited news and Luxembourgish, and link token-level evaluation to downstream writing assistance to quantify the user-facing impact of borrowing misclassifications.

7. Acknowledgements

We thank RTL Luxembourg and Tom Weber for providing access to the news archive and for supporting its use for research purposes. This work

highly benefited from the collaborative network fostered by the **ENEOLI COST Action (CA22126)**, supported by COST (European Cooperation in Science and Technology), and also within the project LuxVoice (project reference 19205922) from the FNR.

8. Ethical and legal aspects

Data provenance and legal basis. The underlying corpus consists of online news articles published between 1999 and 2025 by a major Luxembourgish media outlet (RTL). The data were obtained under a formal research collaboration and processed under the outlet's terms of use and the applicable EU text and data mining provisions for non-commercial scientific research. No user accounts were accessed, no technical protection measures were circumvented, and we did not perform large-scale scraping of the public-facing website.

Data and code availability. Full prompt templates for all strategies and tasks, including the complete five-example few-shot prompt and the neology template, will be provided in the public GitHub: `github.com/NinaKivanani/LexNeo-Bench`.

Intellectual property and data release. All source articles remain under the copyright and database rights of RTL. Our preprocessing, annotation, and analysis operate on copies stored on secure institutional infrastructure; we do not redistribute the full text of the corpus. Instead, we release only derived artifacts that are not substitutable for the original content, including the annotation schema and pattern inventory, scripts to reproduce the pipeline on any legally obtained Luxembourgish news corpus, aggregate statistics and plots, and small illustrative excerpts.

Privacy and data protection. News articles naturally contain references to identifiable individuals. These mentions appear in material already made lawfully available online in the exercise of journalistic freedom, yet they still qualify as personal data. We do not link the corpus to external records, attempt to profile individuals, or infer sensitive attributes. All analyses are conducted at the token, sentence, or aggregate document level rather than at the level of specific persons. Data are stored and processed on secure servers, including national high-performance computing resources, in compliance with the GDPR and relevant national data protection requirements.

Intended use and potential impact. LexNeo-Bench reflects the editorial practices and topic mix of a single news provider and should not be treated as representative of all Luxembourgish language use. The benchmark and LKG are intended as descriptive tools for studying contact phenomena and for stress-testing multilingual LLMs on borrowing and neology, not as prescriptive standards for "correct" Luxembourgish. We explicitly discourage using these resources to police lexical borrowing, to stigmatize code-switching in everyday communication, or to draw strong sociolinguistic conclusions about specific groups. Any correlations between borrowing patterns and social or regional factors must be interpreted with caution to avoid reinforcing stereotypes or over-generalising from a single, institutionally edited source.

9. Bibliographical References

Martine Adda-Decker, Thomas Pellegrini, Eric Bilinski, and Gilles Adda. 2008. Developments of" lëtzebuergesch" resources for automatic speech processing and linguistic studies. In *LREC*.

Elena Alvarez-Mellado. 2020. An annotated corpus of emerging anglicisms in spanish newspaper headlines. In *Proceedings of the 4th workshop on computational approaches to code switching*, pages 1–8.

Elena Álvarez Mellado. 2020. *Lázaro: An extractor of emergent anglicisms in Spanish newswire*. Ph.D. thesis.

Elena Alvarez-Mellado. 2021. Extracting english lexical borrowings from spanish newswire. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 384–386.

Elena Álvarez-Mellado, Jordi Porta-Zamorano, Constantine Lignos, and Julio Gonzalo. 2025. Overview of adobo at iberlef 2025: Automatic detection of anglicisms in spanish. *Procesamiento del Lenguaje Natural*, 75:373–383.

Dimitra Anastasiou. 2022. Deliverable d1. 24 report on the luxembourgish language.

Beiduo Chen, Jun-Yu Ma, Jiajun Qi, Wu Guo, Zhen-Hua Ling, and Quan Liu. 2022. Ustc-nelslip at semeval-2022 task 11: Gazetteer-adapted integration network for multilingual complex named entity recognition. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1613–1622.

Paula Chesley and R Harald Baayen. 2010. Predicting new words from newer words: Lexical borrowings in french. *Linguistics*, 48(6).

Nina Hosseini-Kivanani. 2025. A hybrid framework for neologism validation using llms and lexical knowledge graphs. In *1st International Workshop on Terminological Neologism Management, NeoTerm*, pages 1613–0073.

Nina Hosseini-Kivanani and Fred Philippy. 2026. Luxborrow: From pompier to pompjee, tracing borrowing in luxembourgish. *arXiv preprint arXiv:2603.10789*.

Ran Iwamoto and Hiroshi Kanayama. 2024. Llm neologism: Emergence of mutated characters due to byte encoding. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 24–29.

Laurent Kevers. 2022. Coswid, a code switching identification method suitable for under-resourced languages. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 112–121.

Thomas Lavergne, Gilles Adda, Martine Adda-Decker, and Lori Lamel. 2014. Automatic language identity tagging on word and sentence-level in multilingual text sources: a case-study on luxembourgish. In *LREC*, pages 3300–3304.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 7052–7063.

Elena Álvarez Mellado, Luis Espinosa Anke, Julio Gonzalo Arroyo, Constatine Lignos, and Jordi Porta Zamorano. 2021. Overview of adobo 2021: Automatic detection of unassimilated borrowings in the spanish press. *Procesamiento del Lenguaje Natural*, 67:277–285.

Alistair Plum, Caroline Döhmer, Emilia Milano, Anne-Marie Lutgen, and Christoph Purschke. 2024. Luxbank: The first universal dependency treebank for luxembourgish. In *Proceedings of the 22nd Workshop on Treebanks and Linguistic Theories (TLT 2024)*, pages 30–39.

Zeqi Tan, Shen Huang, Zixia Jia, Jiong Cai, Yinghui Li, Weiming Lu, Yueting Zhuang, Kewei Tu, Pengjun Xie, Fei Huang, et al. 2023. Damonlp at semeval-2023 task 2: A unified retrieval-augmented system for multilingual named entity recognition. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2014–2028.

Aleksandra Tomaszewska, Dariusz Czerski, Bartosz Żuk, and Maciej Ogrodniczuk. 2025. Neon: A tool for automated detection, linguistic and llm-driven analysis of neologisms in polish. In *International Conference on Computational Science*, pages 318–326. Springer.

Jeanine Treffers-Daller. 2025. The simple view of borrowing and code-switching. *International Journal of Bilingualism*, 29(2):347–370.

Sascha Wolfer and Annette Klosa-Kückelhaus. 2023. Tracking the acceptance of neologisms in german: Psycholinguistic factors and their correspondence with corpus-linguistic findings. *Humanities and Social Sciences Communications*, 10(1):1–10.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Jonathan Zheng, Alan Ritter, and Wei Xu. 2024. Neo-bench: Evaluating robustness of large language models with neologisms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13885–13906.

10. Language Resource References

Zenter fir d'Lëtzebuerger Sprooch. 2025. *Lëtzebuerger Online Dictionnaire (LOD)*. Official reference dictionary for Luxembourgish.

11. Appendices

11.1. Supplementary visualization of KG gain

KG gain is defined as the accuracy difference between KG-graph and zero-shot prompting. The gain decreases monotonically with scale, from +56.5 percentage points for Gemma 3 12B to +40.1 for Gemma 3 27B and +36.6 for Llama 3.3 70B.

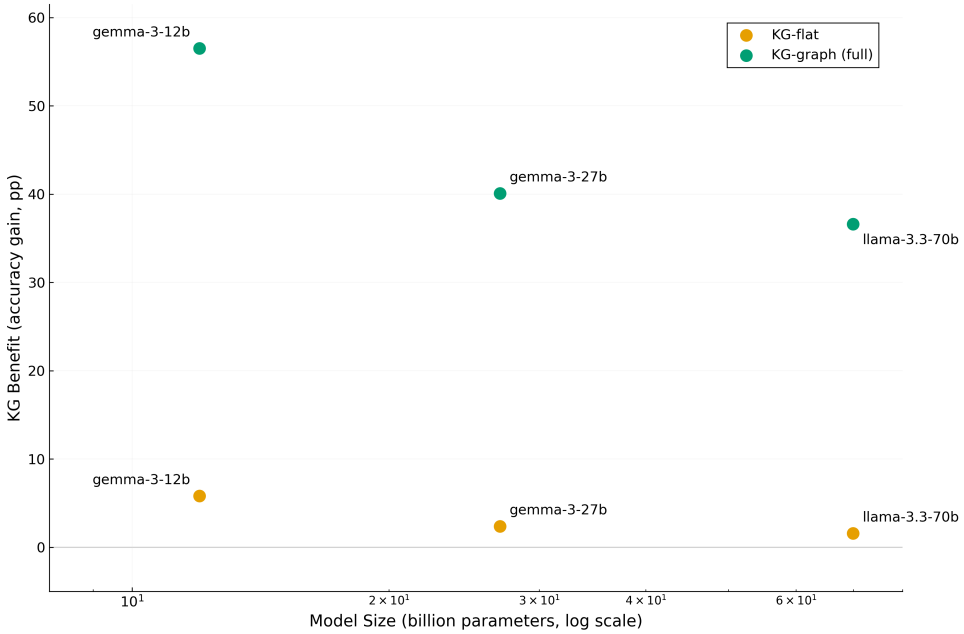

Figure 4: Supplementary visualization of KG gain Δ_{KG} by model size on a log-scaled x-axis.