

Impact of Bilingual Corpus on Pretraining Large Language Models: An Analysis for Turkish

Elif Özge Yılmaz Melike Uysal Çağrı Toraman

Department of Computer Engineering, Middle East Technical University
{yilmaz.ozge_01, melike.uysal, ctoraman}@metu.edu.tr

Abstract

The development of Large Language Models has primarily focused on English-centric or massively multilingual paradigms. As a result, the dynamics of mixing just two languages from the very beginning of training remain underexplored, especially for smaller models with limited capacity. In this study, we conduct a controlled set of experiments to evaluate the impact of bilingual corpus during from-scratch pretraining. Using a capacity-constrained Gemma-4-derived 181M-parameter architecture trained on a 10 billion token budget, we compare a purely Turkish baseline against mixtures containing 10%, 20%, or 30% of an auxiliary language. We test four languages to isolate different linguistic relationships: English (a global bridge language), Azerbaijani (closely related sibling), Finnish (morphological peer), and French (lexical pivot). We show that adding any secondary language costs the model fluency on structurally uniform text (academic writing), a capacity tax that grows with the percentage of auxiliary language. Conversely, every mixture lowers perplexity on stylistically divergent text (creative and literary writing). At matched mixture ratios Azerbaijani yields the lowest perplexity on creative and literary text, ahead of the other auxiliary languages.

1 Introduction

Large Language Models (LLMs) remain overwhelmingly English-centric (Brown et al., 2020). While extending these models to lower-resource languages is a critical research problem, the prevailing paradigm mostly relies on post-training adaptations such as vocabulary expansion or continual pretraining of existing high-resource models. Although computationally efficient, these adaptations inherit the structural habits of the original model and obscure how the two languages actually interact when learned together from scratch. When a

single model is trained on many languages at once, adding more languages eventually starts to hurt performance on all of them, because the model’s capacity stays fixed while the number of languages keeps growing; this effect is known in the field as the curse of multilinguality. Recent large-scale studies indicate that the severity of this issue is influenced by typological proximity, shared scripts, and token mixture ratios (Foroutan et al., 2025; Longpre et al., 2026), but they characterize it through cross-language scaling-law fits or population-level summaries rather than by isolating individual language pairs. Foroutan et al. (2025) report no significant aggregate curse of multilinguality once language token counts are balanced appropriately, and ATLAS (Longpre et al., 2026), while spanning a wide parameter range, estimates transfer at the level of hundreds of language pairs rather than isolating a single typological family under tightly matched conditions. Thus, what remains underexplored is a more specific trade-off within a single fixed-capacity model. Does adding an auxiliary language to the training data cost the model fluency on structurally uniform text, and if so, is that cost offset by gains in robustness on stylistically divergent text, like creative or literary writing? A follow-up question is whether this trade-off tracks typological kinship rather than scale or pivot status. To answer this, we conduct a controlled ablation study using Turkish as a case language (motivated in Section 2.1). We train the same capacity-constrained, 181M-parameter architecture (a scaled-down Gemma-4 design (Gemma Team, Google DeepMind, 2026)) from scratch many times, each run seeing 10 billion tokens total. In the bilingual runs, we replace 10%, 20%, or 30% of the Turkish training data with text from a second, auxiliary language, and train on that mixture from the first step. We call this cross-lingual transfer throughout the paper, but we mean something more specific than the term usually implies. Rather than taking a model al-

ready pretrained on a single language and adapting it to a new one, we are asking what happens when two languages are mixed together from scratch, before the model has fully learned either one. We evaluate four auxiliary pivot languages: English (global pivot), Azerbaijani (genealogically related sibling), Finnish (morphologically similar agglutinative peer), and French (lexical pivot, representing a major source of European loanwords in the Turkish vocabulary). This ablation framework is target-language-agnostic. Turkish is a case study for studying multilingual pretraining dynamics in general, not as an end in itself.

Our main contributions are listed as follows:

- We establish a consistent ranking among four auxiliary languages at matched mixture ratios: Azerbaijani yields the lowest perplexity on stylistically divergent Turkish text, ahead of a morphological peer (Finnish), a lexical pivot (French), and a global pivot (English).
- We provide a systematic empirical analysis of how varying token mixture ratios (10%, 20%, 30%) across four linguistic axes affect from-scratch pretraining dynamics in a capacity-constrained regime. Any auxiliary language degrades fluency on structurally uniform text, monotonically with its share, while the same mixtures reduce perplexity on stylistically divergent text.
- We publicly release all code, data-processing pipelines, and evaluation sets used in this study to support reproducibility¹.

2 Related Work

2.1 Why Turkish as a Case Language

Turkish is a proper testbed for studying cross-lingual transfer. As an agglutinative language with rich derivational and inflectional morphology, it has a very different subword distribution from Indo-European languages, so auxiliary-language contributions cannot be attributed to mere tokenization overlap; its morphology and syntax have long posed challenges for NLP frameworks designed around Indo-European languages (Hakkani-Tür et al., 2000; Oflazer, 2014; Umutlu et al., 2025). It is also a mid-resource language (Joshi et al., 2020),

and has enough clean data for a monolingual baseline, yet is resource-constrained enough that transfer effects remain observable within a 10B-token budget.

Turkish also occupies an unusual position, as it is among the highest-resource Turkic languages, yet remains mid-resource in absolute terms. Turkish and Azerbaijani are in the same language family tree, yet morphologically parallel to unrelated languages such as Finnish. This lets us vary typological kinship, morphological similarity, and lexical overlap across auxiliary languages, which is difficult with a majority-resource language such as English or German. Prior work on Turkish LLM adaptation (Toraman, 2024) shows the challenges of aligning Turkish morphology with English-pretrained architectures, further motivating our from-scratch approach.

2.2 Cross-Lingual Transfer and Typological Proximity

The curse of multilinguality was first formalized by early foundational models like XLM-R (Conneau et al., 2020), which showed that per-language performance in a fixed-capacity multilingual model degrades once the number of co-trained languages passes a certain point. Recent work has both reconfirmed this effect at scale and proposed architectural mitigations for it (Chang et al., 2024; Pfeiffer et al., 2022). This literature characterizes degradation as a function of the number of co-trained languages. We instead hold that number fixed at two and vary the token ratio between them, a distinct manipulation on which the literature is largely silent. Additionally, the Kardeş-NLU benchmark (Senel et al., 2024) showed that using a high-resource (HR) relative such as Turkish significantly improves downstream performance for low-resource (LR) relatives such as Azerbaijani during fine-tuning. Although existing paradigms predominantly explore this unidirectional knowledge flow (HR transferring to LR), our study systematically ablates the source of transfer across four linguistic axes. We investigate the inverse dynamic during from-scratch pretraining, asking whether an LR relative is a more effective auxiliary language than a global pivot for an HR target on stylistically divergent text. This directional asymmetry is consistent with concurrent work on Turkic machine translation (Cinar et al., 2026), which finds that Turkish–Azerbaijani is among the strongest pairwise transfer directions across five Turkic lan-

¹<https://github.com/metunlp/bilingualpretrain>

languages (Turkish, Azerbaijani, Uzbek, Kazakh, Kyrgyz). Their finding is consistent with typological proximity driving transfer strength, though Turkish and Azerbaijani are also the pair with the highest lexical overlap among these languages, so transfer quality in machine translation does not by itself isolate a genealogical mechanism. Concurrent work by Zhang et al. (2026) pretrains bilingual pairs from scratch while varying the input writing system, and finds that romanized pretraining, which maximizes surface-form sharing between languages, yields the strongest cross-lingual transfer. Their result indicates that shared surface form alone can drive transfer, a possibility our design cannot exclude (Section 4.4).

2.3 Data Mixtures and the Capacity Trade-off

The mechanics of pretraining are governed by scaling laws (Kaplan et al., 2020). Large-scale multilingual pretraining spans joint models from bilingual to massively multilingual objectives (Lample and Conneau, 2019; Scao et al., 2022; Shliazhko et al., 2024), none of which isolate the bilingual, from-scratch setting we study here. Additionally, frameworks like ATLAS (Longpre et al., 2026) extend these laws to multilingual environments across a broad range of model sizes (10M–8B parameters) and over 400 training languages. Per-language mixture ratios have also been studied at comparable scale via temperature-based sampling weighting (Xue et al., 2021; Chung et al., 2023), the latter showing that naive heuristics are suboptimal and proposing a fairer allocation. All three operate on the scale of tens to hundreds of languages and report aggregate transfer-scaling or benchmark trends at the population level. Thus, our study is complementary rather than contradictory in scale. Rather than fitting a general-purpose scaling law or sampling policy across many languages, we hold model size, tokenizer, script, and data pipeline fixed and isolate a single auxiliary language at a time under a fixed bilingual mixture ratio. This lets us trace the resulting perplexity change to a specific auxiliary language rather than a single aggregate metric, broken down by evaluation domain (academic texts, creative essays, encyclopedic, literary). Unlike studies evaluating static ratios at large scales, we conduct a fine-grained ablation across 10%, 20%, and 30% auxiliary token mixtures, which (Section 4.2) reveals a capacity trade-off between perplexity on structurally uniform text and perplexity on stylistically divergent text.



Figure 1: Overview of our from-scratch pretraining ablation setup, detailing data sources, token mixture ratios, and evaluation benchmarks.

3 Methodology

An overview of our end-to-end experimental pipeline that illustrates the data sources, our cross-lingual ablation strategy, and the evaluation domains is given in Figure 1.

3.1 Model Architecture

To investigate the effects of cross-lingual transfer, we train our models from scratch using a scaled-down version of the Gemma-4 architecture (Gemma Team, Google DeepMind, 2026). In order to establish a controlled 181M-parameter environment where the effects of a fixed capacity budget can be observed, we configured the architecture with 12 hidden layers, a hidden dimension size of 512, and an intermediate feed-forward size of 2048.

We use a grouped-query attention mechanism with 8 query heads alongside 4 local and 2 global key-value (KV) heads (head dimension of 64). To enhance memory efficiency during sequence processing, a sliding window attention mechanism with a window size of 256 is implemented. Furthermore, we apply a local-to-global attention ratio of 5, meaning a global attention layer is strategically placed after every 5 local layers. To conserve the parameter budget at this scale, the input and output word embeddings are tied.

3.2 Pretraining Data and Tokenization

Each model is pretrained on a corpus of 10 billion tokens. The baseline model is trained exclusively on Turkish data, while the experimental models uti-

lize a bilingual mixture. For the cross-lingual runs, we vary the token mixing ratio across 10%, 20%, and 30% auxiliary language content (with the remainder being Turkish) to evaluate the impact of typological and morphological variations across four auxiliary languages: English, Azerbaijani, French, and Finnish. We chose this range and spacing for two reasons. The range lets us test whether the capacity trade-off scales monotonically or non-monotonically with auxiliary-language share. Also capping the auxiliary share at 30% keeps Turkish the majority language in every configuration, consistent with studying bilingual co-training rather than language replacement.

The Turkish training subset is sampled from the HPLT corpus (de Gibert et al., 2024). Since HPLT already applies document-level deduplication, we did not deduplicate further; we additionally filtered out short, low-quality lines and documents with anomalous or repetitive character distributions, and enforced Turkish language identification with a pretrained fastText classifier (Joulin et al., 2017) to remove non-Turkish or heavily mixed documents.² We applied no perplexity-based quality filtering, to preserve the corpus’s stylistic diversity.

For each auxiliary language, we randomly sample documents up to the token budget required by that configuration’s mixture ratio. Azerbaijani, Finnish, and French are drawn from the corresponding language subsets of FineWeb2 (Penedo et al., 2025) while English is drawn from FineWeb (Penedo et al., 2024), which is English-only. Each auxiliary sample (at most 3B tokens, at the 30% ratio) is drawn without replacement from a fixed 9.6B-token pool per language and seen once during training; no document is repeated. In all cases we apply only the corpus’s default deduplication and length filtering, with no additional per-language quality filtering.

Tokenization across all configurations uses the tokenizer from the google/gemma-4-E2B-it checkpoint (Gemma Team, Google DeepMind, 2026); holding the tokenizer constant ensures observed perplexity differences are attributable to data mixture composition rather than tokenization artifacts, with tokenizer design left as an orthogonal variable for future work.

²Our filtered Turkish subset is available at <https://huggingface.co/datasets/elifozgeyilmaz/turkish-hplt2-filtered>.

3.3 Language Selection and Script Control

Our selection of auxiliary languages was designed to isolate specific linguistic phenomena such as morphological structure and loanword overlap, while controlling for script-level interference as sharing a writing system is itself a major driver of positive cross-lingual transfer (Longpre et al., 2026). Therefore, we restricted the auxiliary languages to the Latin script: Finnish (Ismayilzada et al., 2025), rather than a non-Latin agglutinative language like Japanese, isolates morphological alignment; French (Lewis, 1999), rather than script-mismatched Persian, isolates loanword overlap without the confound of tokenization-level friction. This choice is supported by concurrent evidence that surface-form representation is itself a strong determinant of transfer (Zhang et al., 2026). Holding the script constant therefore removes a variable known to be consequential, though it does not remove lexical overlap within the Latin script, which Section 4.4 discusses.

3.4 Training Infrastructure

All models were trained independently on a single Google Cloud NVIDIA A100 GPU for approximately 60 hours each (10B tokens) under identical hardware and hyperparameter settings; each configuration in Table 1 reflects a single run (see Limitations). In total, we performed 14 such runs, totaling roughly 830 A100 GPU-hours.

3.5 Optimization and Training Details

All models were trained with AdamW (Loshchilov and Hutter, 2019) ($\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.1), a linear warmup over 1,500 steps to a peak learning rate of 3.0×10^{-4} followed by cosine decay to 10% of peak, gradient clipping at a max global norm of 1.0, and bfloat16 mixed precision. Training ran for 153,000 steps with sequences packed/truncated to 1,024 tokens and a per-device batch size of 8 with 8 gradient accumulation steps, corresponding to ≈ 10.03 B tokens per run, on a single device (world size 1).

3.6 Evaluation Setup

We evaluate model robustness and generalization on four Turkish evaluation sets spanning different linguistic styles and structural complexities.

We use standard perplexity as our primary metric, following standard practice in pretraining-scale ablation studies (Kaplan et al., 2020; Chang et al.,

2026; Toraman, 2024), since it requires no task-specific templates and lets us decompose model behavior by evaluation domain rather than a single aggregate score. Perplexity is a known imperfect proxy for downstream performance, particularly in cross-lingual settings and in a typology-specific way (Liang et al., 2023; Yong et al., 2023; Fujinuma et al., 2022; Chang et al., 2026), so we treat our results as evidence rather than a direct proxy for semantic task performance (Section 5). The evaluation sets are as follows:

- **Bellaturca (Academic Split)** (Altinok, 2024): academic texts with the lowest sentence-length variability, clause-punctuation density, and foreign-word ratio of the four sets (Appendix A), making it the most structurally uniform. It is, however, structurally the *least* similar of the four to the HPLT training corpus (Table 3), so we do not attribute its low perplexity (Table 1) to training-distribution overlap (Section 4.2).
- **Essays** (Bilkent Creative Writing Corpus) (Yilmaz, 2025): Represents creative writing and open-ended compositions. Relative to Bellaturca, its distinguishing property is syntactic rather than lexical: Greater sentence-length variety and clausal density at matched vocabulary richness (Appendix A).
- **Turkish Wikipedia** (Wikimedia Foundation, 2026): Represents structured, knowledge-dense encyclopedic text, testing the models’ generalization to formal, objective language.
- **TDK Set (TDK-TD26)**: To evaluate generalization on contamination-free text, we introduce the TDK-TD26 evaluation set, manually curated from the 2026 issues of the Türk Dili periodical (Turkish Language Association (TDK), 2026). Because this content is sourced from a printed journal postdating typical pretraining cutoffs, it is guaranteed to be absent from our web-crawled pretraining corpora. The set contains 20 articles (~38.9K words, 91K Gemma-4 tokens) of literary essays, short stories, and academic criticism, with a 7.4% word-level foreign-word ratio and the highest lexical diversity of the four sets (MATTR = 0.756; full diagnostics in Appendix A).

Throughout the paper, **structurally uniform text** refers to formal expository text with short, regular sentences (Bellaturca), whereas **stylistically divergent text** refers to expressive, non-expository registers whose style is different from our web-crawled pretraining corpus (TDK Set and Essays). Wikipedia TR, formal but structurally more varied falls in neither of those categories and is reported separately.

For each configuration and evaluation set, we estimate uncertainty via a document-level bootstrap (1,000 resamples with replacement), reporting the mean and standard deviation of perplexity across replicates. This captures sampling uncertainty from the finite evaluation set, not variance across independent training runs (see Limitations).

4 Results and Discussion

In this section, we present the empirical evaluation of our from-scratch pretrained models. We report perplexity across all four evaluation sets. Lower perplexity scores indicate superior language modeling performance. Figure 2 visualizes the central qualitative pattern discussed throughout this section (perplexity against auxiliary mixture ratio, one panel per evaluation domain, one line per auxiliary language), and Table 1 reports the full underlying numbers, including per-cell uncertainty.

4.1 Overfitting Control: The 8B Monolingual Run

Before interpreting the transfer results, we verify that the 10B monolingual baseline does not simply reflect overfitting to a saturated training set. We trained an identical architecture on only 8B Turkish tokens and evaluated it on all four evaluation sets: 13.71 ± 0.02 (Bellaturca), 25.18 ± 0.29 (Wikipedia TR), 49.51 ± 0.17 (Essays), 64.23 ± 1.43 (TDK-TD26). All four point estimates are higher (worse) than the corresponding 10B numbers (13.40, 24.63, 48.12, 61.58), consistent with the additional 2B tokens providing generalization gains rather than memorization. The gap is significant on Bellaturca and Essays ($z > 5$), but not on Wikipedia TR or TDK-TD26 ($z = 1.4$ for both); we report the direction on all four but only claim a real 8B-vs-10B difference for the first two. Perplexity on creative text remains comparably severe at 8B (49.51 ± 0.17) as at 10B (48.12 ± 0.17) and both are far higher than any bilingual mixture (26.57–31.44). More monolingual data does not

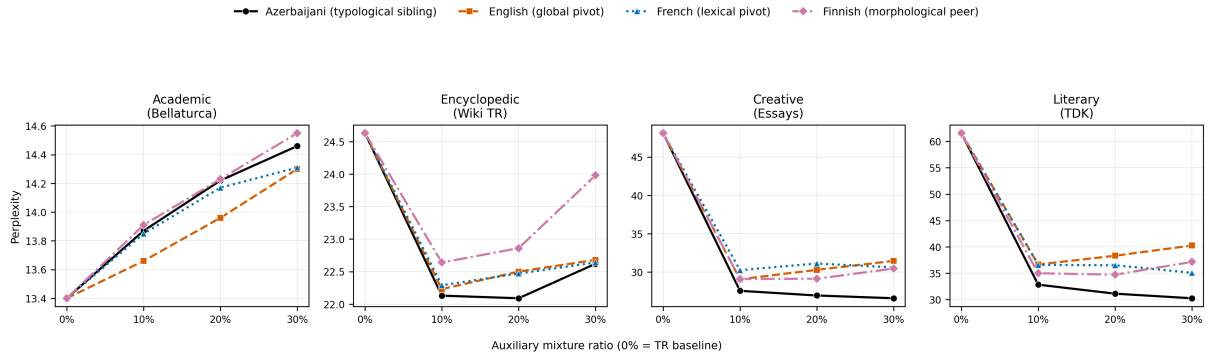


Figure 2: Perplexity vs. auxiliary mixture ratio (0% = TR baseline), one panel per evaluation domain and one line per auxiliary language.

close the gap on creative text. It does not by itself establish why the monolingual condition underperforms here; the HPLT-FineWeb corpus asymmetry (see Limitations) is an alternative contributor we cannot fully rule out with the current design.

4.2 The Capacity Trade-off on Structurally Uniform Text

Table 1 illustrates the perplexity across evaluated domains. The monolingual Turkish baseline achieves the best perplexity on the academic Bellaturca dataset (13.40 ± 0.02). We tested whether this reflects distributional similarity to the HPLT pretraining corpus by directly comparing sentence-structure statistics between HPLT and each evaluation set (Appendix A, Table 3); the comparison does not support that explanation, since Wikipedia TR, not Bellaturca, is structurally closest to HPLT on 3 of 4 measured axes, yet does not achieve the lowest perplexity. We instead attribute the baseline’s strong Bellaturca performance to properties of the register itself: Bellaturca has the shortest mean sentence length and lowest sentence-length variance of the four sets (Appendix A), which plausibly makes local next-token prediction easier independent of any relationship to the training distribution. However, a fine-grained analysis across the 10%, 20%, and 30% token mixture ratios reveals a systematic capacity trade-off. Adding any auxiliary language introduces a capacity tax: a consistent cost to fluency on this structurally uniform text. As the proportion of auxiliary tokens scales, Bellaturca perplexity degrades monotonically for every auxiliary language (e.g., increasing Azerbaijani from 10% to 30% inflates perplexity from 13.87 ± 0.02 to 14.46 ± 0.03). This degradation is consistent with neural scaling laws (Kaplan et al., 2020), under which a tightly bounded parameter budget

allocated in part to a foreign syntax would be expected to cost absolute target-language fluency on standard benchmarks.

4.3 Perplexity on Stylistically Divergent Text

Our main result concerns perplexity on stylistically divergent, expressive text. While the monolingual baseline performs well on the structurally uniform Bellaturca set (Section 4.2), its perplexity on the creative prose of the Bilkent Creative Writing corpus is far higher (48.12 ± 0.17). This gap persists regardless of token budget, as the 8B run confirms, and is specific to the monolingual condition (Table 1). Bilingual co-training is associated with a substantial reduction in perplexity on this stylistically divergent text. Incorporating Azerbaijani changes the model’s perplexity on creative text considerably: even a minimal 10% AZ mixture reduces Creative-domain perplexity from 48.12 ± 0.17 to 27.54 ± 0.09 , and the trajectory is flat to improving at higher ratios, reaching 26.57 ± 0.09 at 30%. Azerbaijani’s advantage coincides with its typological kinship, its morphological similarity, and its lexical overlap with Turkish, three properties that our design cannot separate (Section 4.4). Beyond this correlational observation, we do not have representation-level or causal evidence for a specific mechanism, and leave a mechanistic account for future work. Two tokenizer-level analyses (Appendix B) rule out the most obvious alternative explanations: Azerbaijani has the *highest* tokenizer fertility of all four auxiliary languages, higher even than Turkish itself, ruling out tokenization efficiency; and its word-final suffix overlap with Turkish persists at long suffix lengths where Finnish collapses to zero despite matching or exceeding Turkish’s morphological complexity. Wikipedia TR is the one set on which every bilin-

Model	Ratio	Academic (Bellaturca)	Encyclopedic (Wiki TR)	Creative (Essays)	Literary (TDK)
TR Baseline	100%	13.40 $\pm$ 0.02	24.63 $\pm$ 0.28	48.12 $\pm$ 0.17	61.58 $\pm$ 1.33
Azerbaijani (AZ)	10%	13.87 $\pm$ 0.02	22.13 $\pm$ 0.24	27.54 $\pm$ 0.09	32.82 $\pm$ 0.67
	20%	14.22 $\pm$ 0.03	22.09 $\pm$ 0.24	26.93 $\pm$ 0.09	31.11 $\pm$ 0.62
	30%	14.46 $\pm$ 0.03	22.62 $\pm$ 0.24	26.57 $\pm$ 0.09	30.24 $\pm$ 0.63
English (EN)	10%	13.66 $\pm$ 0.02	22.23 $\pm$ 0.23	29.05 $\pm$ 0.10	36.65 $\pm$ 0.72
	20%	13.96 $\pm$ 0.02	22.50 $\pm$ 0.23	30.28 $\pm$ 0.11	38.31 $\pm$ 0.74
	30%	14.30 $\pm$ 0.03	22.68 $\pm$ 0.23	31.44 $\pm$ 0.11	40.25 $\pm$ 0.82
French (FR)	10%	13.85 $\pm$ 0.02	22.29 $\pm$ 0.23	30.24 $\pm$ 0.11	36.57 $\pm$ 0.70
	20%	14.17 $\pm$ 0.03	22.47 $\pm$ 0.23	31.10 $\pm$ 0.11	36.50 $\pm$ 0.68
	30%	14.31 $\pm$ 0.03	22.64 $\pm$ 0.23	30.60 $\pm$ 0.11	35.07 $\pm$ 0.67
Finnish (FI)	10%	13.91 $\pm$ 0.02	22.64 $\pm$ 0.24	29.08 $\pm$ 0.10	34.99 $\pm$ 0.69
	20%	14.23 $\pm$ 0.03	22.86 $\pm$ 0.24	29.12 $\pm$ 0.10	34.71 $\pm$ 0.67
	30%	14.55 $\pm$ 0.03	23.98 $\pm$ 0.25	30.45 $\pm$ 0.11	37.14 $\pm$ 0.73

Table 1: Perplexity across linguistic domains and mixture ratios; lower is better. Each cell is a single training run. The $\pm$ values are bootstrap standard deviations across 1,000 per-document resamples of the evaluation set; they describe uncertainty from the finite evaluation set and do not bound variation across independent training runs, which we did not measure (see Limitations). **Bold** marks the single best score per column across all 13 configurations. Figure 2 shows the same data as per-domain trend lines.

gual mixture outperforms the monolingual baseline (24.63 $\pm$ 0.28 against 22.09 $\pm$ 0.24 for 20% Azerbaijani) despite being structurally closest to HPLT (Table 3), which a capacity tax alone would not predict. One possibility is that encyclopedic text rewards exposure to a broader range of entities and factual phrasings; another is that the sentence-level metrics in Table 3 do not capture the lexical and factual content that dominates perplexity in this domain. We leave this question to future work.

4.4 Transfer Ranking Across Auxiliary Languages

Our multi-axis ablation shows a consistent transfer ranking among the four auxiliary sources. While the global pivot (English) and morphological peer (Finnish) also improve robustness on stylistically divergent text compared to the baseline, neither matches Azerbaijani’s reduction. Azerbaijani and English diverge in their ratio-scaling dynamics (Figure 2). Azerbaijani’s Creative-domain perplexity is flat to improving from 10% to 30% (27.54 $\pm$ 0.09 $\rightarrow$ 26.57 $\pm$ 0.09), while English moves in the opposite direction, rising at every step (29.05 $\rightarrow$ 30.28 $\rightarrow$ 31.44). The global pivot delivers, at best, a one-time benefit at 10%, which erodes as its share of the budget grows. One hypothesis consistent with these curves is that syntactic divergence displaces target-language capacity as the auxiliary

share grows; we do not measure representations and cannot test this here. French and Finnish are less orderly: French’s Creative-domain trajectory rises then falls, and Finnish’s robustness gain on TDK-TD26 holds through 20% before eroding at 30%. On the Encyclopedic domain (Wikipedia TR), AZ 10% and 20% are indistinguishable at our resolution; morphological and syntactic alignment may aid entity and fact mapping early, with no further gain from higher ratios. French and Finnish both improve over baseline at 10%, though Finnish’s advantage narrows as its share grows and is no longer distinguishable from baseline at 30%. Azerbaijani also has the highest vocabulary overlap with Turkish (Table 6: 0.422 AZ, 0.240 FI, 0.120 FR, 0.100 EN), matching the Creative-domain ranking at 30%. Genealogical proximity and lexical overlap therefore co-vary across our auxiliary languages, and we report the ranking rather than attributing it to either. Kazakh or Uzbek, which are Turkic but lexically more distant from Turkish, would separate the two (Section 5).

4.5 TDK-TD26 Literary Evaluation Set

The TDK-TD26 evaluation set is the most demanding test, as it yields the highest perplexity of any evaluation set under every configuration (Table 1). Structurally, its clause-punctuation density is the highest of the four sets, but its subordinator

rate patterns with Wikipedia TR’s appositive style rather than Essay Corpus’s clausal-argumentative one (Appendix A). Results nonetheless mirror the Creative-domain *ranking* despite this structural difference, which suggests that the advantage is not tied to one particular structural profile. AZ 20% and 30% are indistinguishable at our resolution; Finnish, French, and English follow in that order, all well between Azerbaijani and the monolingual baseline (61.58 ± 1.33). Azerbaijani’s advantage therefore holds across both the Creative and Literary domains. We note two caveats on this set. First, TDK-TD26 contains 20 documents, so the document-level bootstrap resamples from 20 units; its standard deviations are correspondingly wider than those of the other sets and its contrasts should be read more coarsely. Second, TDK-TD26 also has the highest measured lexical diversity of the four sets (MATTR = 0.756, hapax = 0.685; Appendix A), which independently predicts higher perplexity for any language model regardless of typological relationship. With only four evaluation sets we cannot separate this lexical-difficulty effect, or domain difficulty more generally, from a stylistic-robustness effect; all are consistent with the observed numbers.

5 Conclusion and Future Work

In this paper, we investigated from-scratch pretraining dynamics in a controlled, capacity-constrained regime (181M parameters, 10B tokens), using Turkish as a case language. By systematically ablating across four auxiliary languages and three token mixture ratios, we characterized how the choice of auxiliary language and its share of the token budget affect target-language perplexity across registers. On structurally uniform text our results are consistent with the curse of multilinguality where a fixed-capacity model pays for auxiliary-language content whose cost grows monotonically with the auxiliary share for every language we tested. What this framing does not anticipate is that the sign of the effect depends on the evaluation register. On stylistically divergent text the four auxiliary languages separate consistently: at matched ratios Azerbaijani outperforms the morphological peer (Finnish), the lexical pivot (French), and the global pivot (English), and its advantage does not erode as its share grows, while English’s does. We treat this ranking as our central result. The margin between the monolingual baseline and the bilingual

mixtures is larger still, but that comparison also changes corpus source (see Limitations), and we do not rest our claims on it. As discussed previously (Section 3.6), these gains are specifically in robustness to stylistically divergent text, measured by perplexity; we do not claim they equate to improved downstream semantic understanding. In this setting, one target language at 181M parameters on a 10B-token budget, allocating 10–30% of the budget to a typologically close auxiliary language improved perplexity on stylistically divergent text without additional parameters. Whether the same holds at other scales, budgets, or target languages is untested here. Thus, we offer this as evidence from a controlled case study rather than as a general rule for multilingual pretraining.

Future Work. We plan to extend the auxiliary-language axis itself, in order to separate the linguistic properties that co-vary across our auxiliary languages. Contrasting Kazakh and Uzbek against Azerbaijani would do this directly as both are Turkic, but both have substantially lower lexical and subword overlap with Turkish than Azerbaijani does (Table 6). If the typological-sibling advantage survives at low lexical overlap, genealogical proximity is implicated; if it does not, the effect is better described as surface-form sharing. Second, we plan to evaluate on Turkish downstream tasks, including named-entity recognition, natural language inference, and extractive question answering, for example via the Turkish portion of Kardes-NLU (Senel et al., 2024). We scope the present study to perplexity because models trained from scratch at 181M parameters on 10B tokens produce near-floor scores on these benchmarks, which would not distinguish our configurations. Establishing whether the perplexity ranking we report transfers to task performance remains an open question. Third, we plan to extend the model size to a 1B-parameter model trained on a 45B-token budget, to test whether the capacity trade-off and the transfer ranking persist at larger scale. Finally, since the underlying framework is designed to generalize beyond any single target language, we plan to replicate this ablation with target languages outside the Turkic family, to test whether the dynamics we observe reflect a general property of bilingual pretraining rather than an effect specific to Turkish.

Limitations

Several limitations should be considered when interpreting our findings. First, all experiments use a single parameter scale (181M); both the capacity tax and the relative advantage of typological siblings may shift at larger scales. Second, all runs share the same tokenizer (Gemma-4), which was not optimized for Turkish morphology and may affect the magnitude of the observed effects. Third, each configuration reflects a single training run. The uncertainty we report is evaluation-set bootstrap noise and does not bound seed-to-seed training variance, which we have not measured at this scale. Our conclusions therefore rest on effects that are large and consistent across ratios and domains, rather than on individual pairwise contrasts. Thus, small differences, particularly single-step changes, should be read with caution. Multi-seed replication remains future work. Fourth, Turkish data is drawn from the HPLT corpus while all four auxiliary languages come from the FineWeb family (FineWeb2; FineWeb for English); this asymmetry could inflate the TR-baseline-vs-auxiliary gap but, being shared across all four auxiliary languages, it is unlikely to explain their relative ranking. Additionally, a language-identification audit finds negligible Turkish content in the auxiliary corpora (Appendix B). Finally, our evaluation sets differ simultaneously along multiple measured axes (sentence structure, lexical diversity, foreign-word ratio; Appendix A) that are confounded with each other and with typological relationship to Turkish. With only four evaluation sets we cannot attribute perplexity differences to any single one of these factors. We report z -tests only for the 8B-versus-10B comparison in Section 4.1 and do not correct for multiple comparisons. As motivated in Section 3.6, perplexity and downstream performance are known to decouple across cross-lingual configurations, so the structural differences we observe should not be read as established gains on downstream benchmarks. Lastly, Turkish at a 10B-token budget is a mid-resource setting in which auxiliary data is optional; our results do not speak directly to genuinely low-resource regimes where it is not.

Ethical Considerations

Use of Generative AI. Generative AI was used in writing of this study to assist with language editing. All scientific contributions, data construction, data analysis, and interpretations presented in this work

are original and were conducted entirely by the authors.

Data and Model Sourcing. Most corpora used in this study are publicly available resources released by their original authors for research use. The TDK-TD26 evaluation set consists of short excerpts from *Türk Dili*, a journal published by the Turkish Language Association (TDK); copyright for this material belongs to TDK. We use these excerpts only internally in a manner consistent with fair-use norms and we do not release this subset of the data. We do, however, release the metadata needed to reconstruct the set from the source periodical: the article list with titles, authors, issue numbers, years, and page ranges, together with our excerpt-extraction and casefolding procedure, in our code repository. None of the corpora used involve personal, sensitive, or identifying information. Our training and evaluation code and data-processing pipelines are available (see footnote in Section 1). Our filtered Turkish HPLT subset is also available on Hugging Face (Section 3.2). The trained model checkpoints are research artifacts used solely for the comparative analysis reported here and carry no ethical risks beyond those already inherent to the underlying corpora and to language-model pretraining in general.

Acknowledgments

We gratefully thank Google Gemini Academic Reward Program, which helped enable the experiments and computing resources used in this study. Additionally, the numerical calculations reported in this paper were partially performed at TÜBİTAK ULAKBİM, High Performance and Grid Computing Center (TRUBA resources). Lastly, we thank Gökçenaz Akyol, Murat Bayraktar, and Ruşen Birben for the training codebase on which this study builds.

References

- Duygu Altınok. 2024. Bella turca: A large-scale dataset of diverse text sources for turkish language modeling. In *International Conference on Text, Speech, and Dialogue*, pages 196–213, Cham. Springer Nature Switzerland.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1877–1901.

- Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin Bergen. 2026. [Goldfish: Monolingual language models for 350 languages](#). In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pages 3750–3781, Palma de Mallorca, Spain. ELRA Language Resource Association.
- Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2024. [When is multilinguality a curse? language modeling for 250 high- and low-resource languages](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA. Association for Computational Linguistics.
- Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharan Narang, and Orhan Firat. 2023. [UniMax: Fairer and more effective language sampling for large-scale multilingual pretraining](#). In *The Eleventh International Conference on Learning Representations*.
- Omer Burak Cinar, Mehmet Mert Dalkilic, and Cagri Toraman. 2026. Cross-lingual transfer for machine translation in Turkic languages. [ArXiv:2607.29355](#).
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Ona de Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer van der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, Sampo Pyysalo, Stephan Oepen, and Jörg Tiedemann. 2024. [A new massive multilingual dataset for high-performance language technologies](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1116–1128, Torino, Italia. ELRA and ICCL.
- Negar Foroutan, Paul Teilletche, Ayush Kumar Tarun, and Antoine Bosselut. 2025. Revisiting multilingual data mixtures in language model pretraining. [arXiv preprint arXiv:2510.25947](#).
- Yoshinari Fujinuma, Jordan Boyd-Graber, and Katharina Kann. 2022. [Match the script, adapt if multilingual: Analyzing the effect of multilingual pretraining on cross-lingual transferability](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1500–1512, Dublin, Ireland. Association for Computational Linguistics.
- Gemma Team, Google DeepMind. 2026. Gemma 4 technical report. [arXiv preprint arXiv:2607.02770](#).
- Dilek Z. Hakkani-Tür, Kemal Oflazer, and Gökhan Tür. 2000. [Statistical morphological disambiguation for agglutinative languages](#). In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*.
- Mete Ismayilzada, Defne Circi, Jonne Sälevä, Hale Sirin, Abdullatif Köksal, Bhuwan Dhingra, Antoine Bosselut, Duygu Ataman, and Lonneke Van Der Plas. 2025. [Evaluating morphological compositional generalization in large language models](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1270–1305, Albuquerque, New Mexico. Association for Computational Linguistics.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [The state and fate of linguistic diversity and inclusion in the NLP world](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293. Association for Computational Linguistics.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. [Bag of tricks for efficient text classification](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. [arXiv preprint arXiv:2001.08361](#).
- Guillaume Lample and Alexis Conneau. 2019. [Cross-lingual language model pretraining](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Geoffrey Lewis. 1999. *The Turkish Language Reform: A Catastrophic Success*. Oxford Linguistics. Oxford University Press.
- Percy Liang et al. 2023. [Holistic evaluation of language models](#). *Transactions on Machine Learning Research (TMLR)*.
- Shayne Longpre, Sneha Kudugunta, Niklas Muenighoff, I-Hung Hsu, Isaac Rayburn Caswell, Alex Pentland, Sercan O Arik, Chen-Yu Lee, and Sayna Ebrahimi. 2026. [ATLAS: Adaptive transfer scaling laws for multilingual pretraining, finetuning, and decoding the curse of multilinguality](#). In *The Fourteenth International Conference on Learning Representations*.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Kemal Oflazer. 2014. [Turkish and its challenges for language processing](#). *Language Resources and Evaluation*, 48(4):639–653.

- Guilherme Penedo, Hynek Kydlíček, Loubna Ben Al-lal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: decanting the web for the finest text data at scale. In *Proceedings of the 38th International Conference on Neural Information Processing Systems, NIPS '24*, Red Hook, NY, USA. Curran Associates Inc.
- Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. Fineweb2: One pipeline to scale them all – adapting pre-training data processing to every language. *arXiv preprint arXiv:2506.20920*.
- Jonas Pfeiffer, Naman Goyal, Xi Victoria Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. [Lifting the curse of multilinguality by pre-training modular transformers](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States. Association for Computational Linguistics.
- Teven Le Scao et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.
- Lütfi Kerem Senel, Benedikt Ebing, Konul Baghirova, Hinrich Schuetze, and Goran Glavaš. 2024. [Kardeş-NLU: Transfer to low-resource languages with the help of a high-resource cousin – a benchmark and evaluation for Turkic languages](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1672–1688. Association for Computational Linguistics.
- Oleh Shliachko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2024. [mgpt: Few-shot learners go multilingual](#). *Transactions of the Association for Computational Linguistics*, 12:58–79.
- Cagri Toraman. 2024. [Adapting open-source generative large language models for low-resource languages: A case study for Turkish](#). In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 30–44, Miami, Florida, USA. Association for Computational Linguistics.
- Turkish Language Association (TDK). 2026. Türk dili. <https://tdk.gov.tr/yayinlar/sureli-yayinlarimiz/turk-dili-dergisi/>.
- Elif Ecem Umutlu, Ayse Aysu Cengiz, Ahmet Kaan Sever, Seyma Erdem, Burak Aytan, Busra Tufan, Abdullah Topraksoy, Esra Darıcı, and Cagri Toraman. 2025. [Evaluating the quality of benchmark datasets for low-resource languages: A case study on Turkish](#). In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 471–487, Vienna, Austria and virtual meeting. Association for Computational Linguistics.
- Wikimedia Foundation. 2026. Wikimedia downloads: Turkish Wikipedia dump. <https://dumps.wikimedia.org/>. Accessed: 2026-05-31.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. [mT5: A massively multilingual pre-trained text-to-text transformer](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.
- Selim F. Yilmaz. 2025. Compilation of Bilkent Turkish writings dataset. <https://github.com/selimfirat/bilkent-turkish-writings-dataset>.
- Zheng-Xin Yong, Hailey Schoelkopf, Niklas Muenighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Barua, Genta Winata, Stella Biderman, Dragomir Radev, and Vassilina Nikoulina. 2023. [BLOOM+1: Adding language support to BLOOM for zero-shot prompting](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11682–11703, Toronto, Canada. Association for Computational Linguistics.
- Muge Zhang, Aaron Jencks, Krishna Badikela, Yulia Tsvetkov, and Sachin Kumar. 2026. One form to transfer them all: Pretraining multilingual language models beyond native orthography. *arXiv preprint arXiv:2608.25904*.

A Eval-Set Diagnostic Statistics

To characterize our four Turkish evaluation sets on equal footing, we computed three corpus-level diagnostics for each: sentence structure, word-level foreign-language ratio via language identification, and lexical richness. All statistics use the same Turkish-correct casefolding and the same fastText LID model (Joulin et al., 2017) used elsewhere in this work. These analyses are supplementary and not required to interpret Table 1; the paper does not rely on them for its central claims.

Sentence structure. Table 2 reports sentence-length coefficient of variation (CV, scale-free) alongside two parser-free structural proxies: clause-punctuation density (commas, semicolons, colons, dashes per sentence) and subordinator hit rate (occurrences per sentence of a curated list of Turkish subordinating conjunctions, e.g. *ancak* (‘however’), *çünkü* (‘because’), and *rağmen* (‘despite’)). Bel-laturca dataset is the least structurally varied set by

every measure, consistent with its formal, expository register. TDK-TD26 shows the highest clause-punctuation density of all four sets (1.70/sentence) but a subordinator rate close to Wikipedia TR’s rather than Bellaturca’s or Bilkent essays’ which suggests its structural complexity, like Wikipedia’s, is more appositive/parenthetical (citations, asides) than clausal-argumentative.

Evaluation set	CV	Punct./sent.	Subord./sent.
Bellaturca	0.644	0.722	0.480
Essays	0.802	0.851	0.471
Wikipedia TR	0.950	1.136	0.300
TDK-TD26	0.787	1.700	0.305

Table 2: Sentence-length coefficient of variation, clause-punctuation density, and subordinator rate per sentence, across all four evaluation sets.

Comparison to the training corpus (HPLT).

To test whether the sentence-structure differences above reflect similarity to the pretraining distribution, we computed the same analyses on a sample of our cleaned HPLT corpus itself (Table 3). Ranking each evaluation set by absolute distance from HPLT on each of the four metrics and averaging the ranks (1 = closest), Wikipedia TR is structurally closest to HPLT (avg. rank 1.25, closest on 3 of 4 metrics), while Bellaturca is the most structurally divergent from HPLT of the four sets (avg. rank 3.75, farthest on 3 of 4 metrics); Essays and TDK-TD26 fall in between (avg. rank 2.5 each).

Source	Len.	CV	Punct.	Subord.
HPLT (Training)	17.23	1.112	1.175	0.361
Bellaturca	13.62	0.644	0.722	0.480
Essays	14.08	0.802	0.851	0.471
Wikipedia TR	16.14	0.950	1.136	0.300
TDK-TD26	15.10	0.787	1.700	0.305

Table 3: Mean sentence length (words), sentence-length CV, clause-punctuation density, and subordinator rate, for HPLT and each evaluation set.

Foreign-word ratio. Table 4 reports the fraction of eligible words (length ≥ 3 characters) classified as non-Turkish with confidence ≥ 0.5 ; low-confidence calls are excluded from the denominator rather than forced into either class. TDK-TD26’s 7.4% ratio sits between Bellaturca (5.6%) and Wikipedia TR (10.3%).

Lexical richness. Table 5 reports average word length, raw type-token ratio (TTR), Moving-

Evaluation set	Foreign	Uncertain	Top languages
Bellaturca	5.58%	20.2%	en, de
Essays	5.80%	21.3%	en, de
Wikipedia TR	10.26%	29.3%	en, de, fr
TDK-TD26	7.40%	24.1%	en, de, fr

Table 4: Word-level foreign-language ratio (fastText LID, confidence ≥ 0.5) across all four evaluation sets.

Average TTR (MATTR, 500-word window), and hapax legomena ratio. TDK-TD26 has the highest MATTR (0.756) and hapax ratio (0.685) of all four sets, exceeding even the Turkish training-corpus reference (MATTR = 0.718), consistent with its literary-essay register drawing on a wide, non-repetitive vocabulary.

Eval. set	Word len.	TTR	MATTR	Hapax
Bellaturca	6.258	0.068	0.712	0.516
Essays	6.226	0.054	0.712	0.518
Wikipedia TR	6.268	0.138	0.681	0.549
TDK-TD26	6.001	0.372	0.756	0.685

Table 5: Average word length, raw TTR, MATTR (w=500), and hapax legomena ratio across all four evaluation sets.

B Training-Corpus Diagnostics

Section 4.3 notes that Azerbaijani’s robustness advantage coincides with its typological kinship with Turkish. Three alternative explanations are that the effect is instead driven by (a) the Gemma-4 tokenizer happening to represent Azerbaijani more efficiently, (b) generic agglutinative morphological complexity rather than genealogical kinship specifically, or (c) residual Turkish text in the auxiliary samples, which, unlike the Turkish data, were not language-filtered. We computed two tokenizer-level diagnostics on 20M-token samples per auxiliary language (mean $\pm$ std across 3 independent draws, using the same tokenizer as pretraining) to test (a) and (b), and audit the auxiliary corpora for residual Turkish content to test (c).

Tokenizer fertility and vocabulary overlap. Table 6 reports fertility (tokens per word) and top 10,000 token vocabulary overlap with Turkish. Azerbaijani has the *highest* fertility of all four auxiliary languages, which is higher even than Turkish itself. This shows that the tokenizer fragments Azerbaijani more, not less, than English, French, or Finnish.

Lang.	Fertility (tok/word)	Jaccard vs. TR	Overlap coef. vs. TR
TR	2.148 $\pm$ 0.005	—	—
AZ	2.940 $\pm$ 0.009	0.422 $\pm$ 0.008	0.593 $\pm$ 0.007
FI	2.762 $\pm$ 0.026	0.240 $\pm$ 0.005	0.387 $\pm$ 0.007
FR	1.544 $\pm$ 0.005	0.120 $\pm$ 0.002	0.215 $\pm$ 0.003
EN	1.287 $\pm$ 0.001	0.100 $\pm$ 0.0003	0.181 $\pm$ 0.0006

Table 6: Tokenizer fertility and top-10K vocabulary overlap with Turkish, mean $\pm$ std across 3 independent 20M-token draws.

Suffix n-gram overlap. Table 7 reports Jaccard overlap of the top-200 most frequent word-final character n-grams with Turkish, for suffix lengths $n = 2..5$. This tests explanation (b). English, French, and Finnish all collapse to exactly 0.000 overlap by $n = 4-5$, while Azerbaijani gradually decays and never reaches zero (11.8% at $n = 5$). Finnish matches or exceeds Turkish’s morphological complexity by word length and MATTR (Appendix A-style metrics on the training corpus itself), yet still collapses to zero suffix overlap, identical to the typologically unrelated languages. These analyses rule out tokenizer efficiency and generic morphological complexity as explanations; the Turkic-specific suffix overlap is consistent with Azerbaijani’s kinship with Turkish.

Lang.	$n=2$	$n=3$	$n=4$	$n=5$
AZ	0.533 ± 0.010	0.223 ± 0.004	0.148 ± 0.007	0.118 ± 0.004
EN	0.328 ± 0.007	0.026 ± 0.003	0.003 ± 0.000	0.000 ± 0.000
FR	0.330 ± 0.003	0.028 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
FI	0.316 ± 0.004	0.074 ± 0.006	0.003 ± 0.002	0.000 ± 0.000

Table 7: Suffix n-gram Jaccard overlap with Turkish by suffix length, mean $\pm$ std across 3 independent 20M-token draws, top-200 suffixes per length.

Turkish contamination in auxiliary corpora.

Table 8 shows negligible Turkish contamination across all four auxiliary corpora: Finnish, French, and English contain no documents identified as Turkish, while Azerbaijani shows a small but non-zero rate (0.055% of tokens), consistent with its lexical proximity to Turkish (Table 6). This rate is far too low to explain the large perplexity improvements from Azerbaijani co-training (Table 1), so Turkish-language leakage cannot account for its advantage.

Lang.	Docs (N)	Tokens (N)	% Docs as TR	% Tok. as TR
AZ	7,291,231	9.60B	0.0775%	0.0550%
FI	8,365,027	9.60B	0.000%	0.000%
FR	12,179,649	9.60B	0.000%	0.000%
EN	12,894,996	9.60B	0.000%	0.000%

Table 8: Fraction of each auxiliary-language training sample identified as Turkish (`__label__tr`, confidence ≥ 0.80) by the same fastText classifier used to filter the Turkish HPLT corpus.