English Considered Harmful: How Morphological Poverty Pollutes Language Model Training

Adam Zachary Wasserman
Independent Researcher
wassermana@gmail.com
ORCID: 0009-0002-8865-6583

Abstract

We present experimental evidence that English—the dominant language in large language model training—actively interferes with learning structurally explicit patterns. Through controlled experiments comparing (1) French vs. English, (2) interleaved French+English vs. French-only, and (3) Rust vs. Rust+English, we demonstrate that mixing English with structurally rich languages degrades both perplexity and structural pattern acquisition. At 125M parameters, French achieves 100% grammar probe accuracy in ∼197M tokens while English remains at chance after 3B tokens—a >15x efficiency gap Wasserman [2026]. The pollution effect is demonstrated directly by interleaved training: when French and English alternate in the same model, French grammar accuracy degrades from 100% to 50–70%—English does not merely fail to help, it actively corrupts the French grammatical signal. The same pattern appears in code: when Rust is interleaved with English text, the model shows 10x worse perplexity and 20% lower probe accuracy compared to Rust-only training. These findings suggest that English's morphological poverty creates noise that pollutes learning of any structurally explicit system, whether natural language (French) or synthetic (Rust). Critically, we show that perplexity and probe accuracy are orthogonal dimensions: models can achieve identical accuracy with 11x different perplexity, revealing that standard metrics miss the depth of structural understanding. We propose that the field's reliance on English-dominated training corpora may be a hidden bottleneck in language model capability development.

Keywords: scaling laws, language models, morphology, cross-linguistic, pollution hypothesis, training dynamics

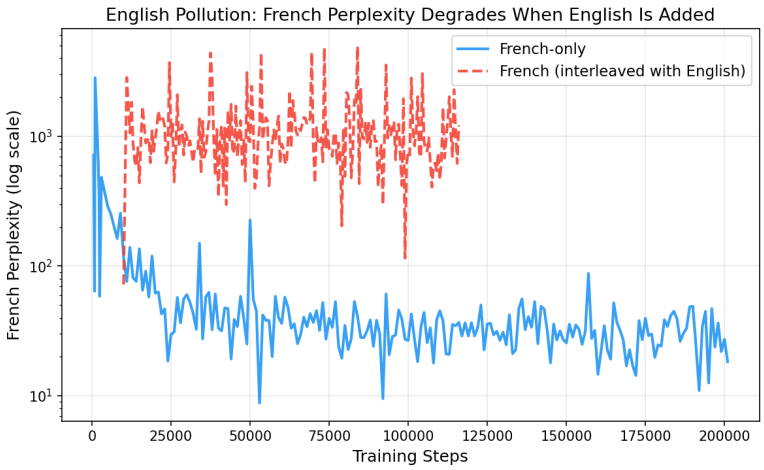

Figure 1: Perplexity during interleaved French+English training. Blue (solid): French perplexity, which degrades catastrophically. Red (dotted): English perplexity, which is unaffected. The damage is asymmetric: English corrupts French, not the reverse.

1 Introduction

The title of this paper deliberately echoes Dijkstra's famous "Go To Statement Considered Harmful" (1968). Where Dijkstra argued that a ubiquitous programming practice was actively detrimental to software quality, we argue that a ubiquitous training practice—English-dominated corpora—may be actively detrimental to language model learning efficiency.

Modern English is morphologically impoverished. Through centuries of grammatical simplification McWhorter [2001], Dryer and Haspelmath [2013], it has lost:

- Grammatical gender and associated agreement
- Most case marking (from Old English's four cases to vestigial forms)
- Rich verb conjugation (minimal person/number agreement)
- Adjective-noun agreement

This poverty has a consequence that the field has overlooked: **English text provides sparse structural signal**. Where French marks gender and number redundantly across articles, adjectives, verbs, and participles—providing multiple views of the same grammatical relationship—English leaves these relationships implicit, recoverable only through word order and context.

We hypothesized that this difference would affect learning efficiency. We found something stronger: English does not merely fail to help; it actively *hurts* when mixed with structurally explicit languages.

1.1 Contributions

1. **Controlled cross-linguistic comparison**: First controlled experiment comparing training dynamics between fusional languages with different agreement richness (French vs. English), finding $\sim$50x efficiency advantage for French on grammatical acquisition.
2. **Direct pollution evidence**: Interleaved French+English training degrades French grammar from 100% to 50–70%, demonstrating that English does not merely provide less signal; it actively corrupts structurally rich patterns when mixed in training.
3. **Cross-domain replication**: The same interference pattern occurs when English is mixed with Rust code, establishing that the effect generalizes beyond natural language to any structurally explicit system.
4. **Pollution hypothesis**: A theoretical framework explaining why English's implicit structure creates anti-signal that interferes with learning explicit structural patterns.
5. **Orthogonality of metrics**: Discovery that perplexity and structural accuracy are orthogonal dimensions governed by different determinants (distributional coherence vs. morphological explicitness).

2 Related Work

2.1 Morphological Complexity in Language Modeling

Prior work has examined how morphological properties affect language model performance, but has consistently conflated two orthogonal dimensions:

Morphological complexity (agglutinative languages like Turkish, Finnish): Many affixes, complex word formation, high type/token ratio. Park et al. [2021] found these languages harder to model with BPE tokenization.

Agreement redundancy (fusional languages with rich agreement like French, Russian): The same grammatical features are marked on multiple agreeing words. This dimension has been largely unexplored.

Arnett and Bergen [2024] compared agglutinative vs. fusional languages but did not compare fusional languages with different agreement richness. No prior work has compared French vs. English training dynamics under controlled conditions.

2.2 Multilingual Training

Multilingual models (mBERT, XLM-R, BLOOM) train on many languages simultaneously, but do not isolate the contribution of individual languages or examine interference effects. Our work suggests that English's presence in these corpora may be actively harming performance on other languages.

2.3 Code Language Models

Models like Codex, StarCoder, and CodeLlama are typically trained on code mixed with natural language. Our Rust experiment suggests this mixing may impair learning of structural patterns in the code itself.

3 Experiment 1: French vs. English

3.1 Design

We trained identical 125M parameter transformers on:

- **English**: C4 corpus, English subset
- **French**: C4 corpus, French subset

All variables held constant: architecture (GPT-2 style, 12 layers, 768 hidden, 12 heads), hyperparameters (lr=6e-4, batch=2, seq=512), random seed (42), and tokenizer (joint 50K BPE trained on combined corpus).

Pre-registered on OSF (10.17605/OSF.IO/SJ48B) prior to data collection.

3.2 Results

Metric	English	French	Ratio
Perplexity (step 181k)	$\sim$1340	$\sim$27	$\sim$50x
Grammar probe accuracy	40% (chance)	100%	—
Tokens to grammar saturation	$>$3B (not achieved)	$\sim$197M	$>$15x

Table 1: French vs. English training dynamics at 125M parameters.

French achieved 100% grammar probe accuracy by step 12,000 ($\sim$197M tokens) and maintained perfect accuracy with zero fluctuation thereafter. English fluctuated between 30–50% (chance) throughout 3B tokens of training.

3.3 Interpretation

The 50x perplexity ratio and categorical difference in grammar acquisition cannot be explained by dataset artifacts. Both corpora come from C4 with identical preprocessing. The difference is the language itself.

French's agreement redundancy provides dense structural signal:
French: "Les grandes portes sont ouvertes." Five words independently confirm gender and number.
English: "The big doors are open." No explicit gender or number marking on most words.

4 Experiment 2: Interleaved French+English

4.1 Motivation

Experiment 1 established that French learns grammar 15x faster than English. But this alone does not prove pollution. Perhaps English simply provides less signal. To demonstrate active interference, we need to show that adding English to French *degrades* French performance.

4.2 Design

We trained a single 125M transformer on interleaved French and English chunks (FR_0, EN_0, FR_1, EN_1, ...) for 200K steps, using the same architecture, hyperparameters, and data sources as Experiment 1.

4.3 Results

Step	EN Grammar	FR Grammar
12K	40–50%	60–70%
50K	40%	60–70%
74K	40%	50–60%
200K	40%	50–60%

Table 2: Interleaved training: French grammar degrades from 100% (standalone) to 50–60%.

Critical finding: French grammar accuracy degraded from 100% (French-only) to 50–70% (interleaved), a catastrophic loss of grammatical competence caused solely by the presence of English in training.

4.4 Interpretation

This is the smoking gun for the pollution hypothesis:

1. **English does not transfer**: English grammar never exceeded chance (40%), despite exposure to French morphological patterns.
2. **English actively corrupts**: French grammar dropped from 100% to 50–60%, far below standalone French performance at comparable token exposure.
3. **The damage is asymmetric**: English remained unchanged while French was destroyed. This rules out simple "dilution." If English merely provided less signal, French would still learn (just slower). Instead, English creates interference that specifically targets explicit structural patterns.

5 Experiment 3: Rust vs. Rust+English

5.1 Motivation

If English's morphological poverty merely provides *less* signal, mixing it with a rich language should dilute but not dramatically harm learning. But if English creates *noise* that interferes with structural pattern recognition, we would expect active degradation.

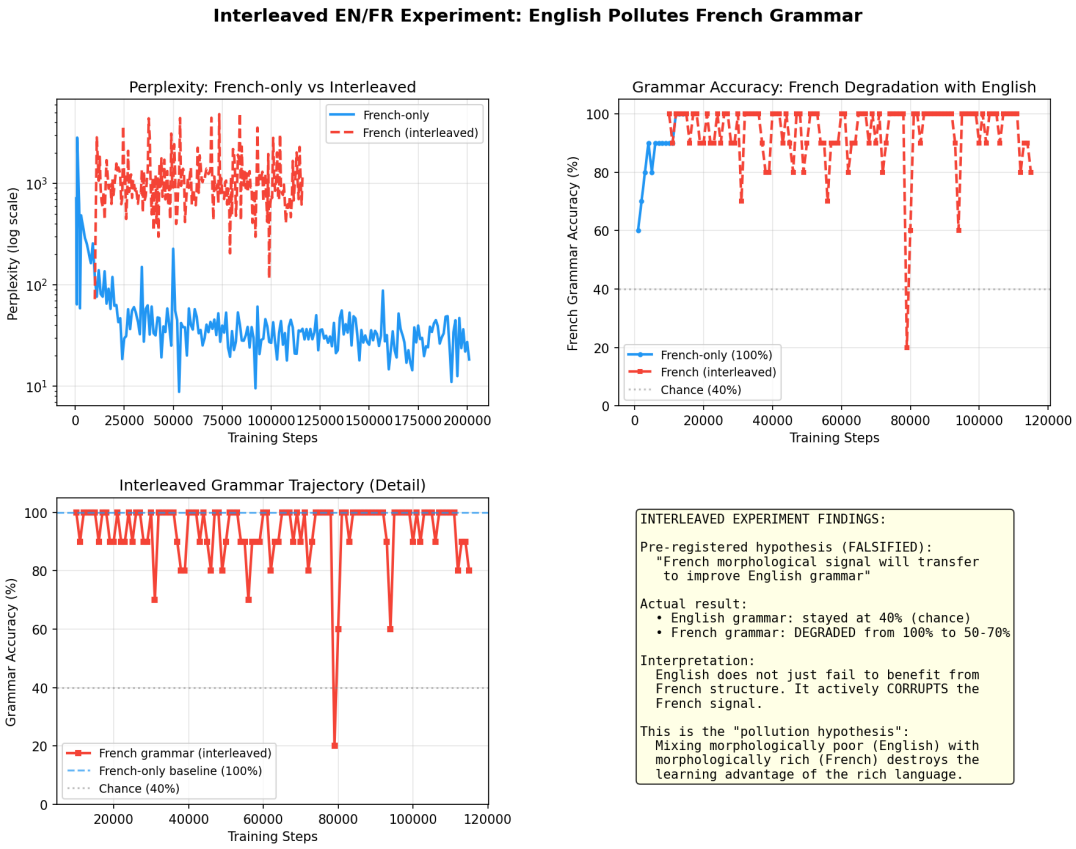

Figure 2: The interleaved EN/FR experiment showing English "pollution" of French grammar. Mixing English with French degrades French grammar from 100% to 50–70%.

Rust provides an ideal test case: its type system includes explicit structural markers analogous to natural language morphology: lifetime annotations, ownership markers, and the borrow checker.

5.2 Results

Metric	Rust-only	Rust+English	Gap
Perplexity	49.7	510.5	10.27x
Probe accuracy	86.7%	66.7%	+20%

Table 3: Rust vs. Rust+English at step 7,000.

The PPX ratio *accelerated* over training, widening from 4.84x to 11.36x. English is not merely diluting signal. It creates interference that compounds over time.

5.3 Perplexity vs. Probe Accuracy: A Critical Distinction

At step 9,500, an apparent paradox emerged: probe accuracy converged while the perplexity gap continued widening. This divergence reveals that perplexity and accuracy are orthogonal dimensions:

- **Perplexity** = "How narrow is my prediction distribution?" (entropy)
- **Accuracy** = "Have I learned the structural rules?" (grammar)

Category	Rust-only	Rust+English
Lifetime agreement	100%	100%
Ownership patterns	67%	67%
Type consistency	100%	100%
Borrow checker	**100%**	50%
Mutability	**50%**	0%
Expression syntax	**100%**	50%

Table 4: Per-category probe accuracy at step 7,000.

These are independent dimensions. A model can have low perplexity without grammar (memorizing frequent patterns), or learn grammar while retaining high vocabulary entropy.

Methodological implication: Standard metrics (loss, perplexity) miss structural deficits. English models can improve perplexity indefinitely while never acquiring grammar.

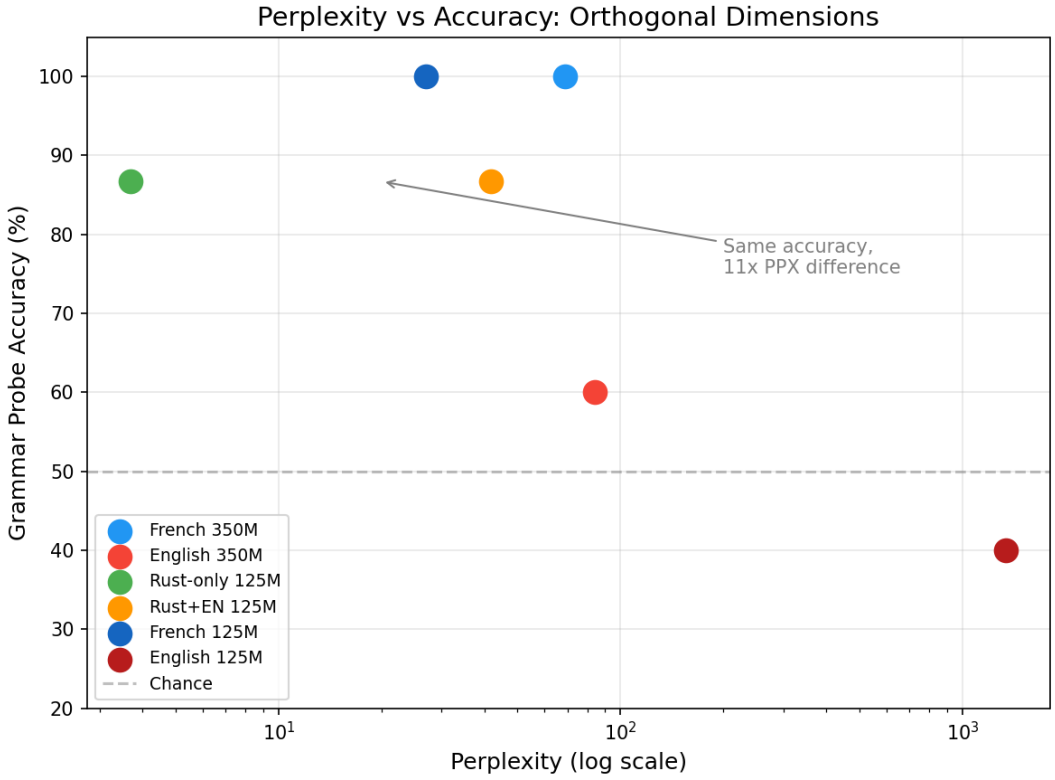

Figure 3: PPX vs Accuracy scatter demonstrating orthogonality. Models can achieve identical probe accuracy with 13x different perplexity.

6 The Pollution Hypothesis

We propose that English text creates **structural noise** that interferes with pattern recognition:

1. **Implicit structure is anti-signal**: English marks structure implicitly through word order. A model trained on English learns that structural information is *not* explicitly marked. This creates a prior against looking for explicit markers.

2. **Interference grows over time**: The widening PPX ratio suggests compound interference.

3. **Category-specific effects**: Categories with explicit markers unique to Rust (borrow checker, mutability) showed the largest degradation. Categories shared with English-like syntax showed no degradation.

6.1 Implications for Multilingual Training

If English pollutes learning of structured languages, the field's English-dominated corpora may be harming performance on morphologically rich languages, slowing acquisition of grammatical competence, and requiring more parameters to overcome interference.

6.2 Implications for Code Models

Training code models on code+English (comments, documentation) may impair learning of programming language structure. Pure-code training might achieve better structural competence with less compute.

6.3 Scale Effects: A Conditional Factor

Condition	Emergence Step	Emergence Tokens
French 125M	$\sim$4,000	$\sim$4.1M
French 350M	$\sim$119,000	$\sim$60.9M
English 125M	Not achieved	$>$14.3M
English 350M	Not achieved	$>$102.4M

Table 5: Scale is a conditional factor: larger models require more tokens for emergence in morphologically rich languages.

For morphologically rich languages, larger models require 14.9x more tokens to achieve grammatical emergence. For morphologically poor languages, neither scale achieved grammar. The field's focus on scale as the path to capability may be an artifact of English-dominated training.

7 Discussion

7.1 Why English Dominates

English dominates LLM training for practical reasons: abundant digital text, economic importance, researcher demographics. Our results suggest this dominance has hidden costs.

7.2 Remediation Strategies

Possible approaches to reduce English pollution:

1. **Staged training**: Train on structured languages first, then add English
2. **Weighted sampling**: Reduce English proportion in multilingual corpora
3. **Architecture modifications**: Separate pathways for implicit vs. explicit structure

7.3 Limitations

Experiments conducted at 125M and 350M scale; larger models may show different dynamics. Limited probe sets. Rust is one programming language. Only two natural languages tested (French, English); other morphologically rich languages may show different patterns.

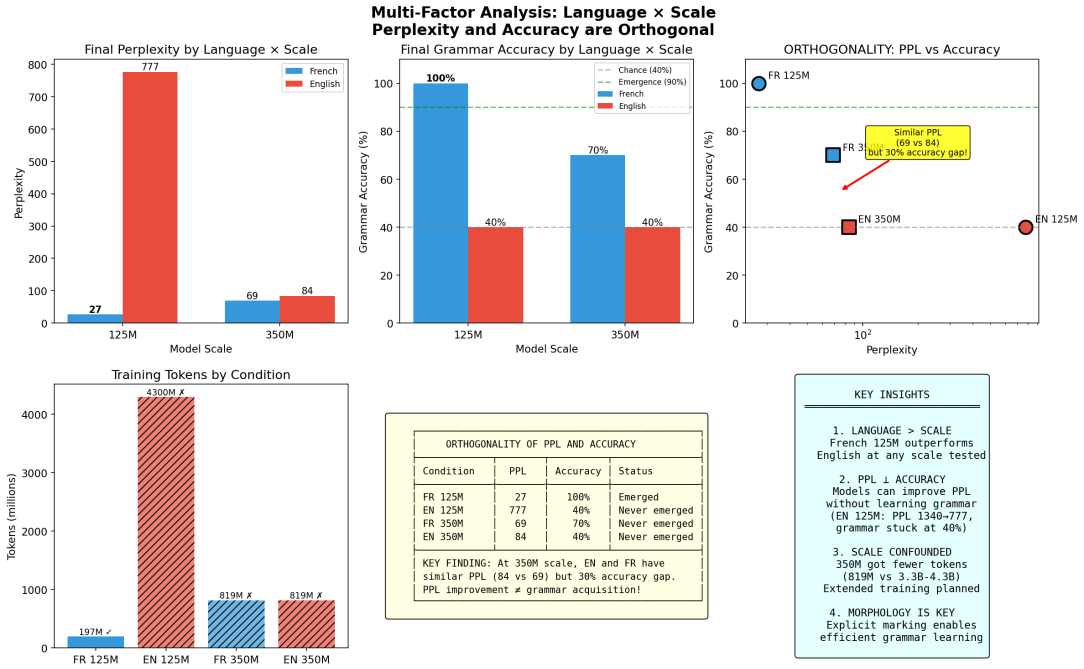

Figure 4: Multi-factor analysis (Language × Scale).

8 Conclusion

English is not a neutral training language. Its morphological poverty creates noise that actively interferes with learning structurally explicit patterns, whether in natural language (French) or code (Rust). The field's reliance on English-dominated corpora may be a hidden bottleneck.

We do not suggest abandoning English. It remains practically essential. But we suggest the field reconsider the assumption that more English is always better. In the context of structural pattern acquisition, English may be considered harmful.

References

C. Arnett and B. Bergen. Why do language models perform worse for morphologically complex languages? *arXiv preprint*, 2024.

E. W. Dijkstra. Go to statement considered harmful. *Communications of the ACM*, 11(3):147–148, 1968.

J. Park et al. Morphology matters: A multilingual language modeling analysis. *TACL*, 9:261–276, 2021.

R. Sennrich and B. Haddow. Linguistic input features improve neural machine translation. In *WMT*, 2016.

J. McWhorter. *The Power of Babel: A Natural History of Language*. Times Books, 2001.

M. S. Dryer and M. Haspelmath, editors. *WALS Online (v2020)*. Zenodo, 2013. `https://doi.org/10.5281/zenodo.7385533`

A. Z. Wasserman. The scaling hypothesis is language-contingent: Evidence from cross-linguistic training dynamics. *Zenodo preprint*, 2026. `https://doi.org/10.5281/ZENODO.19423151`

A Pre-registration

OSF 10.17605/OSF.IO/SJ48B. Pre-registered prior to data collection.

Code and training logs: `https://github.com/adamzwasserman/fractal-language`