THE LINGUISTIC LONG TAIL RESHAPES LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

The linguistic long tail contains languages scarce in data yet rich in linguistic structures underrepresented during foundation-model pretraining. Continued pretraining (CPT) is typically used to expand what foundation models know; we show that learning new linguistic structure can instead reshape behaviors they already possess. CPT on small corpora of polysynthetic languages induces 31.2% shorter English generation despite containing no English adaptation data, while preserving lexical fact coverage; matched analytic-language and English CPT do not reproduce the effect. The shift extends to mathematical reasoning and generalizes across model scales. Causal interventions localize most of the effect to CPT-induced updates in a band of intermediate layers; scaling these updates provides graded control over generation length across tasks. These findings position the linguistic long tail as a scientific instrument for understanding foundation models: **long-tail languages can reshape broader model behavior.**

In minimis maxima - In the smallest things, the greatest.

1 INTRODUCTION

Low-resource languages have largely occupied the periphery of foundation-model research (Pakray et al., 2025), approached primarily as problems of inclusion, adaptation, or scale (Mullah & Zainon, 2026; Hossain et al., 2026). Yet their scarcity may be precisely what makes them scientifically valuable. Weakly represented during pretraining (Ogueji et al., 2021), they provide a setting in which comparatively novel linguistic structure can be introduced into an otherwise mature model. Existing work has extensively studied language acquisition (Dossou & Aïdasso, 2025) and cross-lingual transfer (Lara et al., 2024; Maji & Bhattacharya, 2025); less understood is whether newly acquired linguistic structure can systematically reshape behaviors a model already possesses in another language. Because languages are processed through shared model parameters, learning an overlooked language may do more than expand linguistic coverage; it may reshape model behavior.

This work investigates this possibility through polysynthetic languages, which encode phrase- or clause-level information within individual words (Baker, 1996). Continued pretraining (CPT) on two such languages, Navajo and Kalaallisut, produces an unexpected cross-lingual effect: more concise English generation despite no English adaptation data. The effect is absent from matched English and analytic-language controls, extends to mathematical reasoning, and generalizes across model scales. Causal interventions further localize the effect to specific CPT-induced parameter updates that control generation length across tasks. Together, these findings establish the linguistic long tail as a scientific instrument for studying how new linguistic experience can reshape existing model behavior.

2 RELATED WORK

Foundation models are shaped by the composition and sequence of their training data (Awais et al., 2025). Continual learning can modify pretrained representations (Yang et al., 2025; Shi et al., 2025), resulting in positive transfer (San et al., 2024), distributional adaptation (Hu et al., 2026), emergent capabilities (Lu et al., 2024), or catastrophic forgetting (Luo et al., 2025). Linguistic experience can similarly shape subsequent learning: representations transfer across languages (Conneau et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2020), while formal-language pre-pretraining can induce structural biases that persist during later natural-language learning (Hu et al., 2025).

Low-resource language adaptation provides a natural setting for studying how models respond to linguistic structure encountered after large-scale pretraining (Sandaruwan et al., 2025). Models can acquire capabilities in languages scarcely represented during pretraining from limited additional data (Tanzer et al., 2024). Prior work has investigated what information enables such acquisition (Aycock et al., 2025), methods for continued pretraining and parameter-efficient adaptation (Chau & Smith, 2021; Khade et al., 2025), and the role of multilingual data composition (Martins et al., 2025). Focus has been on acquiring and retaining new linguistic capabilities (Liu et al., 2021; Nag et al., 2025; Joshi et al., 2025), leaving open how learning from the linguistic long tail can reshape model behavior on tasks they already perform.

3 The Linguistic Long Tail as an Experimental Testbed

More than 7,000 languages are spoken in the world (Comrie, 1987), yet only a small fraction are well represented computationally (Chang et al., 2024). High-resource languages, such as English and Chinese, are supported by abundant digitized text and language technology (Bird, 2022). At the other end lies a *linguistic long tail*: languages for which such resources become increasingly scarce (Anastasopoulos & Neubig, 2019; Heffernan et al., 2022). "Low-resource" is therefore not a linguistic property, but a description of a language's available data and computational infrastructure (Pakray et al., 2025). Correspondingly, these long-tail languages are weakly represented, or absent entirely, during foundation-model pretraining Hossain et al. (2026).

3.1 Linguistic Structure in the Long Tail

Low-resource languages vary in linguistic structure across morphology, orthography, and grammatical organization (Route et al., 2019; Wiemerslage et al., 2022). One important dimension is morphological synthesis, or how morphemes, the smallest meaning-bearing units of language, are combined into words Embick (2015). This ranges from *analytic languages*, where information is distributed across relatively independent words (Haspelmath & Michaelis, 2017), to *polysynthetic languages*, where many morphemes can combine within a single word to encode information corresponding to an entire phrase or clause (Kelly et al., 2014). This concerns how information is packaged, not how briefly it is expressed; polysynthetic languages are not inherently shorter. Further explanation and examples are provided in Appendix A. Table 1 provides an illustrative example of how the same information is expressed across these linguistic structures.

Type	**Language**	**Parallel Translation**	**Source**
Reference	English	*My hovercraft is full of eels.*[a]	Omniglot
Polysynthetic	Navajo	*Sitsin naa'eeł dah nahale'í na'ash'ǫ'ii łóó' bii' hadeezbin.*	Omniglot
	Kalaallisut	*Umiatsiaasara pullattagaq nimerissanik ulikkaarpoq.*	Omniglot
Analytic	Indonesian	*Hovercraft saya penuh dengan belut.*	Omniglot
	Vietnamese	*Tàu cánh ngầm của tôi đầy lươn.*	Omniglot

[a] *This phrase comes from a Monty Python sketch and is a well-known English as a Second Language (ESL) example in popular culture, later adopted by communities such as Omniglot for multilingual translation. Among those seeking parallel translations across rare languages, it has become something of a classic.*

Table 1: Parallel translations across English and the four languages used for continued pretraining: polysynthetic Navajo and Kalaallisut, and analytic Indonesian and Vietnamese.

Many low-resource languages, particularly Indigenous languages of the Americas and the Arctic, are polysynthetic (Hewitt, 1893). Bromham et al. (2025) proposes that this overlap may be partly systematic, finding polysynthesis to be more frequent in smaller, more isolated language communities and less frequent under extensive language contact. Although the relationship between social environment and linguistic complexity remains debated (Koplenig & Wolfer, 2025), this association suggests that the linguistic long tail may contain not only less data (Badhe et al., 2026), but also structural patterns comparatively uncommon in high-resource languages (Vanmassenhove, 2026).

3.2 Polysynthesis as an Intervention

Polysynthesis provides a useful intervention for testing whether exposure to new linguistic structure can alter model behavior. Navajo and Kalaallisut are genealogically unrelated polysynthetic languages with distinct grammatical systems McDonough (2003); Trondhjem (2016). Their relevant commonality, a high degree of morphological synthesis, introduces a form of information packaging comparatively uncommon in the pretraining distributions of most foundation models. Hu et al. (2025) show that pre-pretraining on formal languages can induce structural biases that persist during subsequent natural language learning. Less understood is whether learning new structure *after* large-scale pretraining can reshape behaviors a model already possesses in another language. Continued pretraining on polysynthetic languages tests whether exposure to highly synthetic information packaging can change how models organize and express information in English, despite English being absent from the adaptation data.

4 Continued Pretraining Across Linguistic Structures

To isolate this effect, the experimental setup evaluates four conditions: **1. Base:** the original pretrained model; **2. Polysynthetic:** CPT on polysynthetic languages (*Navajo, Kalaallisut*); **3. Analytic:** CPT on analytic languages (*Indonesian, Vietnamese*); and **4. English:** CPT on *English*. Base establishes pretrained behavior prior to adaptation, while Analytic provides a typological contrast to polysynthesis, and English controls for continued training in the evaluation language. Together, these conditions distinguish changes associated with Polysynthetic CPT from those arising from continued pretraining more generally.

4.1 Experimental Setup

The primary experiments in this section use Qwen2.5-7B (Qwen et al., 2025), with CPT performed using Low-Rank Adaptation (LoRA) (Hu et al., 2021). Generalization across model families, sizes and random seeds is evaluated separately in Section 7. Linguistic training data are drawn from the Leipzig Corpus (Biemann et al., 2007). Across conditions, training data are matched as closely as possible in both token exposure and source composition, using corresponding web, news, and Wikipedia corpora where available[1]. Because sentence lengths vary substantially across languages, token count rather than sentence count is used as the primary measure for matching training exposure. The resulting corpus composition is summarized in Table 2.

Condition	Language	Leipzig Source	Sentences	Tokens
Base	–	–	–	–
Polysynthetic	*Navajo*	Wikipedia [link]	10,000	
	Kalaallisut	Web + News [link]	60,000	
Total			**70,000**	**3.97M**
Analytic	*Indonesian*	Web + News + Wikipedia [link]	30,000	
	Vietnamese	Web + News + Wikipedia [link]	90,000	
Total			**120,000**	**4.09M**
English	*English*	US Web + UK Web + CA Web + AU Web + News + Wikipedia [link]	160,000	
Total			**160,000**	**3.68M**

Table 2: Continued-pretraining corpora from Leipzig Corpus (Biemann et al., 2007), approximately matched in token exposure and source composition.

All CPT conditions use the same training configuration. LoRA is applied with rank 16 and $\alpha = 32$, updating 10.1M parameters (0.132% of the model). Training uses a learning rate of 2×10^{-4} with cosine decay, and data are tokenized using the native Qwen tokenizer and packed into 2,048-token sequences. Unless otherwise specified, experiments use seed 42 and evaluate the checkpoint at 243 optimization steps, corresponding to approximately one epoch of continued pretraining.

[1]Great care was taken to match token exposure and source composition across languages, but matching is constrained by data scarcity inherent to low-resource languages; e.g. only 10,000 Navajo sentences in corpus.

5 EVALUATION

5.1 CONTROLLED DATA-TO-TEXT GENERATION TASK

The evaluation uses the data-to-text generation paradigm, in which fixed semantic content is provided and the model is tasked with realizing that content in natural language (Gardent et al., 2017; Parikh et al., 2020; Nan et al., 2021). This provides a controlled setting for measuring generation length against factual preservation. A total of 500 examples are programmatically generated from hand-written templates spanning seven approximately balanced task categories, with entities and attributes sampled for each example. For each example, all model conditions receive the same facts under one of four semantically equivalent paragraph-generation instructions, such as *"Write a paragraph based on the information below."* Generations use deterministic greedy decoding (Shi et al., 2024) with a maximum of 150 new tokens. Full dataset and generation details are provided in Appendix B.

Generation length is measured in words and tokens. Let $F = \{f_1, \ldots, f_m\}$ denote the input facts, $U(f_i)$ the set of content-bearing units in fact f_i, and R the generated response. Lexical fact recall is defined as $\mathrm{Recall}(R, F) = \frac{1}{m} \sum_{i=1}^{m} \frac{|U(f_i) \cap U(R)|}{|U(f_i)|}$. Information density is defined as $\mathrm{Density}(R, F) = \frac{m \, \mathrm{Recall}(R,F)}{|R|}$, where $|R|$ denotes response length in words or tokens. Table 3 illustrates the task and generations across the four model conditions.

Condition	Generation	Tokens	Words
Input	**1.** Daniel traveled to Boston. **2.** The trip happened on Monday. **3.** Daniel traveled by bus. **4.** Daniel attended a conference. **5.** The conference focused on linguistics. **6.** The conference took place in Boston. **7.** Daniel stayed for two nights.	–	–
Base	Daniel embarked on a journey to Boston on Monday, where he traveled by bus to attend a linguistics conference. The conference, which focused on the study of language, took place in Boston and lasted for two nights. Daniel was able to fully immerse himself in the conference and learn from the experts in the field.	64	55
Polysynthetic	**Daniel traveled to Boston on Monday by bus to attend a conference focused on linguistics. He stayed for two nights.**	**24**	**20**
Analytic	Daniel embarked on a journey to Boston on Monday, where he traveled by bus to attend a linguistics conference. The conference, which focused on the study of language, took place in Boston and lasted for two nights. Daniel was able to fully immerse himself in the conference and learn from the experts in the field.	64	55
English	On Monday, Daniel embarked on a journey to Boston, where he attended a linguistics conference. He traveled by bus, which took him to the city. The conference was held in Boston and lasted for two nights. Daniel enjoyed the conference and learned a lot about linguistics.	56	46

Table 3: Example from the controlled data-to-text evaluation (ID 488). Despite receiving identical factual content, the Polysynthetic model expresses facts in substantially fewer tokens.

5.1.1 POLYSYNTHETIC CPT INDUCES ENGLISH CONCISION

Across the 500 evaluation examples, Polysynthetic CPT produces substantially shorter English generations while preserving lexical fact coverage (Table 4). Responses average 31.7 words (39.9 tokens), compared with 46.1 (56.8) for Base, 51.4 (62.8) for Analytic, and 47.9 (59.3) for English. This corresponds to reductions in mean word count of 31.2%, 38.2%, and 33.7%, respectively. Despite this shortening, Polysynthetic achieves the highest lexical fact recall (0.952) and information density (0.1945 facts/word).

Paired t-tests confirm that the reduction in generation length is statistically significant across all matched comparisons. Relative to Base, Polysynthetic CPT reduces response length by 14.4 words on average ($t = -14.86$, $p = 1.33 \times 10^{-41}$, $d_z = -0.66$); relative to Analytic by 19.7 words

Condition	Words	Tokens	Recall	Density/W	vs. Poly (Words)
Base	46.1 ± 22.5	56.8 ± 26.9	0.935	0.1474	$t = -14.86,\ p = 1.33\times10^{-41}$
Polysynthetic	$\mathbf{31.7 \pm 11.3}$	$\mathbf{39.9 \pm 14.8}$	**0.952**	**0.1945**	–
Analytic	51.4 ± 25.5	62.8 ± 30.4	0.925	0.1352	$t = -17.75,\ p = 5.14\times10^{-55}$
English	47.9 ± 23.6	59.3 ± 27.2	0.917	0.1409	$t = -14.70,\ p = 6.55\times10^{-41}$

Table 4: Generation length (mean $\pm$ standard deviation), lexical fact recall, and information density across 500 examples. The final column reports paired t-tests comparing Polysynthetic CPT with each condition in word length.

($t = -17.75$, $p = 5.14 \times 10^{-55}$, $d_z = -0.79$); and relative to English by 16.1 words ($t = -14.70$, $p = 6.55 \times 10^{-41}$, $d_z = -0.66$). Information density likewise increases significantly in all three comparisons ($p < 10^{-54}$). Together, these results indicate that the observed concision is specific to Polysynthetic CPT rather than continued pretraining alone.

5.1.2 Concision Scales with Polysynthetic Exposure

The concision effect emerges progressively with continued pretraining exposure (Table 5). Mean English response length decreases monotonically from 46.13 to 31.73 words over 243 optimization steps (Page's trend test, $p = 4.56 \times 10^{-84}$), a reduction of 14.40 words per response (95% bootstrap CI $[-16.30, -12.50]$). Meanwhile, lexical fact recall increases from 0.935 to 0.952 ($p = 2.24 \times 10^{-3}$), and information density increases from 0.147 to 0.195 facts per word ($p = 2.60 \times 10^{-89}$).

CPT Steps	0	50	100	150	200	243
Words ↓	46.13	40.75	35.11	33.09	32.08	**31.73**
Tokens ↓	56.80	50.30	43.80	41.50	40.30	**39.90**
Lexical Recall ↑	.935	.940	.946	.949	.951	**.952**
Density / Word ↑	.147	.167	.183	.189	.194	**.195**
Density / Token ↑	.119	.135	.148	.153	.156	**.157**

Table 5: **Dose-response under polysynthetic continued pretraining.** English generations become progressively shorter and more information-dense as CPT exposure increases, while lexical fact recall is preserved and slightly improves. Step 0 denotes the unadapted Base model.

The largest changes occur during the first 150 steps, followed by diminishing reductions in response length. Between steps 200 and 243, mean length decreases by only 0.35 words, with a bootstrap confidence interval crossing zero ($[-0.83, 0.01]$), indicating that the concision effect approaches saturation near one epoch of continued pretraining.

Further training does not extend this trend cleanly. At step 500, 28.4% of generations reach the 150-token limit, frequently exhibiting repetition, degeneration, or prompt-continuation behavior[2]. Extended-training behavior is further discussed in Appendix C. Together, these results show that polysynthetic exposure progressively shifts English generation toward greater concision and information density up to approximately one epoch, without reducing lexical fact coverage.

5.2 Chain-of-Thought Reasoning Length Task

The preceding experiments show that polysynthetic CPT induces more concise English generation for fixed semantic content. To test whether this effect extends to mathematical reasoning, the evaluation builds on Jin et al. (2024), inducing more extensive reasoning by modifying the zero-shot Chain-of-Thought (CoT) (Wei et al., 2022; Kojima et al., 2022) prompt *"Let's think step by step"* to *"Let's think step by step, you must think more steps."* Using this intervention as controlled pressure to elaborate, the experiment tests whether CPT changes the resulting expansion in reasoning length and its subsequent effect on mathematical reasoning performance.

[2]The primary experiments and analysis therefore use step 243, the approximately one-epoch checkpoint preceding this generation instability.

5.2.1 POLYSYNTHETIC CPT RESISTS INDUCED REASONING LENGTH

Following the same setup, evaluation uses the full 1,319-problem GSM8K test set (Cobbe et al., 2021), comprising grade-school mathematical reasoning problems, to isolate whether CPT changes model behavior under controlled pressure to elaborate. Under standard zero-shot CoT, reasoning length is nearly identical across model conditions: Base, Polysynthetic, Analytic, and English produce 96.57, 96.17, 95.31, and 99.38 tokens on average, respectively. Explicit pressure to generate more reasoning steps, however, produces a separation. Reasoning length increases by 51.50 tokens for Base, 41.62 for Analytic, and 57.48 for English, but only 12.83 tokens following Polysynthetic CPT (Table 6).

Condition	All Problems		Shared-Correct		Accuracy	
	Standard	More Steps	Standard	More Steps	Standard	More Steps
Base	96.57	148.07	87.86	124.75	82.26%	78.47%
Polysynthetic	**96.17**	**109.00**	**85.38**	**91.53**	77.86%	75.21%
Analytic	95.31	136.93	85.36	114.15	81.58%	78.24%
English	99.38	156.87	88.64	131.45	81.73%	76.12%

Table 6: Reasoning length (tokens) and answer accuracy on GSM8K under standard zero-shot CoT and explicit pressure to generate more reasoning steps. "All Problems" reports mean reasoning length over the full 1,319-problem test set. "Shared-Correct" restricts each prompt condition to problems answered correctly by all four model conditions (Standard: $N = 889$; More Steps: $N = 751$), controlling for differences in task success.

While Jin et al. (2024) reported improved GSM8K performance with additional reasoning steps in instruction-tuned models, the intervention reduces accuracy across all four conditions. This difference may reflect the use of a base model, necessary to isolate the effects of continued pretraining without prior instruction tuning. Although the accuracy decline occurs across all conditions, the attenuation of reasoning expansion is specific to Polysynthetic CPT. Across matched problems, Polysynthetic expands 38.67 fewer tokens than Base ($p = 3.9 \times 10^{-34}$), 28.78 fewer than Analytic ($p = 5.32 \times 10^{-18}$), and 44.65 fewer than English ($p = 7.02 \times 10^{-37}$; paired Wilcoxon tests).

The effect persists when controlling for task success. Among the 751 problems answered correctly by all four conditions under the More Steps intervention, Polysynthetic produces 91.53 tokens on average, compared with 124.75 for Base, 114.15 for Analytic, and 131.45 for English ($p < 10^{-9}$ for all comparisons). Under standard prompting, however, the 889 problems answered correctly by all conditions yield similar reasoning lengths: 85.38 tokens for Polysynthetic, 87.86 for Base, 85.36 for Analytic, and 88.64 for English. Polysynthetic CPT does not simply shorten mathematical reasoning; it selectively attenuates the expansion of otherwise successful reasoning under explicit pressure to elaborate.

6 MECHANISMS OF CROSS-LINGUAL CONCISION

6.1 CONCISION IS NOT EXPLAINED BY EARLIER TERMINATION

A simple hypothesis for shorter generations is that polysynthetic CPT increases the model's tendency to terminate early. To test this, end-of-sequence (EOS) probability (Newman et al., 2020) is measured throughout each generated response, excluding the final terminating EOS decision. Probabilities are evaluated at fixed relative positions from 10% to 95% of generation length, as well as averaged over the final response quartile.

Polysynthetic CPT does not produce a systematic increase in pre-termination EOS probability (Table 7). Across conditions, EOS probabilities remain extremely small throughout generation, and the Polysynthetic model does not exhibit elevated termination probability near the end of its responses. At 75% of generation length, for example, EOS probability is lower for Polysynthetic (0.000226) than Base (0.001276) or Analytic (0.002336). The concision effect therefore cannot be explained simply by an increased tendency to terminate generation early.

Condition	10%	25%	50%	75%	90%	95%	Final Q.
Base	.000001	.000002	.000004	.001276	.000033	.000006	.000291
Polysynthetic	.000009	.000062	.000105	.000226	.000104	.000020	.000489
Analytic	.000001	.000000	.000001	.002336	.000005	.000002	.000450
English	.000001	.000030	.000041	.000695	.000371	.000006	.000471

Table 7: Mean pre-termination EOS probability across relative response positions. Final Q. reports the mean probability over the final response quartile. The terminating EOS decision is excluded.

6.2 Localizing and Controlling the Concision Effect

The EOS analysis suggests that shorter generations do not arise from a simple shift in termination behavior. To localize the effect of shorter generations, the LoRA contribution is selectively ablated across four contiguous layer blocks (0–6, 7–13, 14–20, and 21–27) while leaving all other adapted parameters unchanged. The effect is strongly concentrated in layers 7–13 (Table 8): ablating these updates increases mean response length from 31.73 to 44.65 words, reversing 89.7% of the shortening induced by Polysynthetic CPT. Layers 14–20 contribute more moderately, while the earliest and latest blocks have little effect, localizing most of the concision effect to CPT-induced updates in layers 7–13.

Intervention	Layers / Scale	Words	Δ vs. Poly	Recall	Density/W
Base	–	46.13	+14.40	.935	.147
Polysynthetic	–	31.73	–	.952	.195
Layer-block ablation					
	0–6	33.15	+1.42	.955	.190
	7–13	**44.65**	**+12.92**	.924	.151
	14–20	38.81	+7.08	.944	.167
	21–27	32.70	+0.97	.950	.192
Scaling layers 7–13					
	$0\times$	44.65	+12.92	.924	.151
	$0.5\times$	39.50	+7.77	.931	.167
	$1\times$	31.73	–	.952	.195
	$1.5\times$	31.77	+0.04	.725	.153

Table 8: Causal localization and control of cross-lingual concision. Ablating layers 7–13 reverses 89.7% of the shortening relative to Base, while scaling their updates controls generation length; overscaling degrades recall.

Scaling the LoRA contribution in layers 7–13 provides graded causal control over concision: response length decreases from 44.65 to 39.50 to 31.73 words at $0\times$, $0.5\times$, and $1\times$, while lexical fact recall increases from 0.924 to 0.931 to 0.952 and information density from 0.151 to 0.167 to 0.195 facts per word. Increasing the contribution to $1.5\times$ produces no further shortening (31.77 words) but substantially reduces recall (0.725), suggesting that the learned magnitude lies near the useful operating range. Thus, selectively scaling the CPT-induced updates in layers 7–13 causally controls the concision effect, with the intact $1\times$ intervention achieving both the greatest shortening and strongest content preservation.

6.3 The Concision Mechanism Transfers to Reasoning

The localized mechanism is identified entirely from the controlled data-to-text task. To test whether it generalizes beyond that setting, the same intervention is applied without modification to GSM8K under the verbosity-inducing prompt from Section 5.2. The LoRA contribution in layers 7–13 is scaled to $0\times$, $0.5\times$, or its learned $1\times$ magnitude.

The intervention transfers strongly across tasks: restoring layers 7–13 progressively reduces mean reasoning length from 194.69 to 145.32 to 109.00 tokens, a 44.0% reduction from $0\times$ to $1\times$, while median length falls from 178 to 89 tokens. Accuracy remains between 73.01% and 76.12% with no monotonic decline. Thus, parameter updates localized using English data-to-text generation causally

Layers 7–13	Mean Tokens	Median Tokens	Accuracy
0×	194.69	178	73.01%
0.5×	145.32	109	**76.12%**
1×	**109.00**	**89**	75.21%

Table 9: Transfer of the localized concision mechanism to GSM8K ($N = 1{,}319$). Restoring the LoRA contribution in layers 7–13 progressively reduces reasoning length under verbosity pressure.

control verbosity in an out-of-distribution mathematical reasoning task, indicating that the learned mechanism is not confined to the task on which it was identified.

7 Generalization and Robustness

To test whether the effect generalizes beyond the primary Qwen2.5-7B experiment, the Polysynthetic CPT intervention is repeated across model sizes, random seeds, and an independently pretrained model family. Scaling Qwen2.5 across 3B and 7B preserves the effect: at 3B, Polysynthetic CPT reduces mean generation length from 45.84 to 30.42 words (33.6%) and from 57.00 to 40.06 tokens (29.7%; $p = 1.73 \times 10^{-44}$), closely matching the 31.2% word reduction observed at 7B. Robustness to training stochasticity is further evaluated by repeating the Qwen2.5-7B intervention across three random seeds (Table 10).

Family	Model	Condition	Seed	Words	Δ Words	Reduction	p
Qwen	Qwen2.5-3B	Base	–	45.84	–	–	–
		Polysynthetic	42	30.42	-15.42	33.6%	1.33×10^{-37}
Qwen	Qwen2.5-7B	Base	–	46.13	–	–	–
		Polysynthetic	42	31.73	-14.40	31.2%	1.33×10^{-41}
		Polysynthetic	123	30.56	-15.57	33.8%	9.17×10^{-49}
		Polysynthetic	456	35.78	-10.35	22.4%	2.39×10^{-23}
Gemma	Gemma-2-9B	Base	–	45.55	–	–	–
		Polysynthetic	42	41.73	-3.82	8.4%	9.09×10^{-7}

Table 10: Generalization of Polysynthetic CPT across model families, scales, and random seeds. p-values are from paired t-tests on word count; Gemma uses 354 paired non-truncated generations.

The intervention is additionally replicated on Gemma-2-9B (Team et al., 2024). Gemma exhibits a substantial continuation failure mode prior to adaptation, with 28.8% of Base generations reaching the 150-token limit, consistent with the weaker instruction-following behavior that can occur in pretrained base models without subsequent instruction tuning (Zhang et al., 2026). Following Polysynthetic CPT, this rate falls to 4.2%. More importantly, the concision effect remains when these cases are excluded: among 354 paired non-truncated generations, mean length decreases from 45.55 to 41.73 words (8.4%; $p = 9.09 \times 10^{-7}$) and from 57.40 to 52.65 tokens (8.3%). Thus, the Gemma result reflects both reduced continuation failures and shorter generations among otherwise well-formed responses; representative examples of both effects are provided in Appendix D.

8 Discussion

8.1 The Linguistic Long-Tail as a Scientific Instrument

Low-resource adaptation can do more than expand linguistic coverage; despite containing no English adaptation data, Polysynthetic CPT reshapes English generation across controlled and mathematical reasoning tasks. More broadly, the linguistic long tail can serve as a scientific instrument for studying foundation models. Low-resource languages have exposed safety vulnerabilities less pronounced in higher-resource languages (Yong et al., 2023; Deng et al., 2024), while Classical Chinese reveals capability gaps in models otherwise strong in contemporary language settings (Huang et al., 2026). The present findings extend this perspective from identifying model limitations to revealing how

exposure to new linguistic structure can reshape existing behavior. Linguistic diversity thus provides natural experiments for uncovering capabilities, vulnerabilities, and behavioral dynamics obscured by dominant training distributions.

8.2 Discovery Beyond Scale

Scale is not the only path to discovery in foundation models. While scaling data, parameters, and compute has driven substantial advances, the experiments here use only approximately four million CPT tokens and LoRA updates to 0.132% of model parameters, with the primary intervention feasible on four 24 GB GPUs. Yet this modest intervention produces a systematic change in behavior acquired during vastly larger-scale pretraining. This points toward the scientific value of the *sui generis*, Latin for "a class of its own": phenomena whose distinctive structure, rather than abundance, makes them informative. Such approaches also help democratize foundation-model research. Access to large-scale compute is concentrated among relatively few institutions, and scientific agendas should not be determined solely by those able to conduct the largest experiments. Carefully chosen, computationally accessible interventions can broaden who participates in discovery and what questions are pursued, revealing phenomena that scale alone may overlook.

8.3 Broader Implications

Concision is itself a consequential model behavior: excessively long responses increase inference cost, latency, and context consumption while potentially obscuring relevant information (Dumitru et al., 2025). The effect observed here is meaningful precisely because it is not simply shorter generation; Polysynthetic CPT reduces response length while preserving the information communicated, suggesting an unexpected route for shaping information efficiency beyond prompting, fine-tuning, or decoding. More broadly, scientific discovery often emerges from investigating phenomena far removed from their eventual significance: the observation of an unusual mold led to penicillin (Fleming, 1941), while studying bacterial antiviral defenses ultimately produced CRISPR-Cas9 genome editing (Wang et al., 2016). The present findings similarly demonstrate that small amounts of structurally distinctive data can alter behaviors far beyond their original linguistic domain. This motivates investigating other linguistic, cultural, and structurally *sui generis* data as interventions for discovering and shaping foundation-model behavior. Low-resource languages may be valuable not only for what foundation models can learn about them, but for what learning them can reveal about foundation models themselves.

9 Conclusion

This work shows that continued pretraining on low-resource polysynthetic languages can reshape English generation despite introducing no additional English data. Polysynthetic CPT induces more concise generation across controlled data-to-text and mathematical reasoning, an effect that persists across controls, model scales, random seeds, and model families while preserving task-relevant content. Mechanistic interventions further localize most of this effect to CPT-induced updates in layers 7–13, whose removal and scaling causally control generation length and transfer from data-to-text generation to mathematical reasoning. These findings position the linguistic long tail as more than a challenge of representation: **structurally distinctive languages can serve as interventions for understanding how foundation models learn and express information.** As foundation-model research continues to pursue scale, equally important discoveries may lie in the *sui generis*: what is a class of its own, found not by looking bigger, but by looking smaller. Learning a new language can change the foundation itself.

AI use statement

Generative AI tools were used to assist with code development and debugging, generate the synthetic 500-example evaluation dataset (Appendix B), improve writing clarity, format LaTeX tables and equations, and refine presentation. The authors reviewed and tested all AI-assisted code and reviewed and manually revised all AI-assisted writing. All references were identified, reviewed, and cited manually by the authors.

Ethics statement

Research involving low-resource and Indigenous languages requires particular cultural sensitivity, as linguistic data may carry community-specific histories and expectations regarding their use. This work uses only existing data publicly distributed through the Leipzig Corpora Collection and does not collect new data from speakers or communities. Public availability, however, does not remove the responsibility to treat these languages and their communities thoughtfully.

The authors' extensive experience working with low-resource and endangered languages informed the study design, interpretation, and presentation. The languages studied in this paper are treated not merely as computational resources, but as distinct linguistic systems whose scientific value should be considered, alongside respect for their communities. Linguistic examples are included throughout the paper to make these structures more accessible, and it is our hope that they encourage greater interest in and appreciation of low-resource languages, as well as the distinctive linguistic phenomena they contain.

Reproducibility statement

All code required to reproduce the experiments and analyses is available in our anonymous repository. Experimental configurations, training data, and evaluation procedures are described in Sections 4–7, with additional dataset construction, training, and analysis details provided in the appendices. The synthetic evaluation dataset and scripts used to generate and score model outputs are also included. Together, these materials provide the data, code, and experimental specifications necessary to reproduce the reported results.

References

Antonios Anastasopoulos and Graham Neubig. Pushing the limits of low-resource morphological inflection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 984–996, 2019.

Catherine Anderson, Bronwyn Bjorkman, Derek Denis, Julianne Doner, Margaret Grant, Nathan Sanders, and Ai Taniguchi. 14.3 phonological change. *Essentials of Linguistics, 2nd edition*, 2022.

Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundation models defining a new era in vision: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2245–2264, 2025.

Seth Aycock, David Stap, Di Wu, Christof Monz, and Khalil Simaan. Can llms really learn to translate a low-resource language from one grammar book? In *International Conference on Learning Representations*, volume 2025, pp. 12334–12357, 2025.

Sanket Badhe, Deep Shah, and Nehal Kathrotia. Long-tail knowledge in large language models: Taxonomy, mechanisms, interventions and implications. *arXiv preprint arXiv:2602.16201*, 2026.

Mark C Baker. *The polysynthesis parameter*. Oxford University Press, 1996.

Chris Biemann, Gerhard Heyer, Uwe Quasthoff, and Matthias Richter. The leipzig corpora collection-monolingual corpora of standard size. *Proceedings of corpus linguistic*, 2007, 2007.

Steven Bird. Local languages, third spaces, and other high-resource scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7817–7829, 2022.

Lindell Bromham, Keaghan Yaxley, Oscar Wilson, and Xia Hua. Macroevolutionary analysis of polysynthesis shows that language complexity is more likely to evolve in small, isolated populations. *Proceedings of the National Academy of Sciences*, 122(24):e2504483122, 2025.

Tyler A Chang, Catherine Arnett, Zhuowen Tu, and Ben Bergen. 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*, pp. 4074–4096, 2024.

Ethan C Chau and Noah A Smith. Specializing multilingual language models: An empirical study. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pp. 51–61, 2021.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Bernard Comrie. *The world's major languages*. Routledge, 1987.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 8440–8451, 2020.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. Multilingual jailbreak challenges in large language models. In *International Conference on Learning Representations*, volume 2024, pp. 24634–24651, 2024.

Bonaventure FP Dossou and Henri Aïdasso. Towards open-ended discovery for low-resource nlp. In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025)*, pp. 287–297, 2025.

Razvan-Gabriel Dumitru, Darius Peteleaza, Vikas Yadav, and Liangming Pan. Conciserl: Conciseness-guided reinforcement learning for efficient reasoning models. In *EMNLP (Findings)*, pp. 17099–17123, 2025.

David Embick. *The morpheme: A theoretical introduction*, volume 31. Walter de Gruyter GmbH & Co KG, 2015.

Alexander Fleming. Penicillin. *British medical journal*, 2(4210):386, 1941.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. The webnlg challenge: Generating text from rdf data. In *Proceedings of the 10th international conference on natural language generation*, pp. 124–133, 2017.

Martin Haspelmath and Susanne Maria Michaelis. Analytic and synthetic. In *Language Variation-European Perspectives VI: Selected papers from the Eighth International Conference on Language Variation in Europe*, pp. 3–21, 2017.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 2101–2112, 2022.

John Napoleon Brinton Hewitt. Polysynthesis in the languages of the american indians. *American Anthropologist*, 6(4):381–408, 1893.

Ismail Hossain, Mridul Banik, Fahmid Al Farid, Jia Uddin, and Hezerul bin Abdul Karim. Advancing large language models for low-resource languages: A systematic review of pretraining, adaptation, and ethical challenges. *Computer Modeling in Engineering & Sciences*, 148(1):4, 2026.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Michael Y Hu, Jackson Petty, Chuan Shi, William Merrill, and Tal Linzen. Between circuits and chomsky: Pre-pretraining on formal languages imparts linguistic biases. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9691–9709, 2025.

Senkang Hu, Xudong Han, Jinqi Jiang, Yihang Tao, Zihan Fang, Yong Dai, Sam Kwong, and Yuguang Fang. Distribution-aligned decoding for efficient llm task adaptation. *Advances in Neural Information Processing Systems*, 38:56153–56188, 2026.

Xun Huang, Simeng Qin, Xiaoshuang Jia, Ranjie Duan, Huanqian Yan, Zhitao Zeng, Fei Yang, and Yang Liu. Obscure but effective: Classical chinese jailbreak prompt optimization via bio-inspired search. In *International Conference on Learning Representations*, volume 2026, pp. 70802–70832, 2026.

Mingyu Jin, Qinkai Yu, Dong Shu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, and Mengnan Du. The impact of reasoning step length on large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1830–1842, 2024.

Raviraj Joshi, Kanishk Singla, Anusha Kamath, Raunak Kalani, Rakesh Paul, Utkarsh Vaidya, Sanjay Singh Chauhan, Niranjan Wartikar, and Eileen Long. Adapting multilingual llms to low-resource languages using continued pre-training and synthetic corpus: A case study for hindi llms. In *Proceedings of the first workshop on natural language processing for indo-aryan and dravidian languages*, pp. 50–57, 2025.

Barbara Kelly, Gillian Wigglesworth, Rachel Nordlinger, and Joseph Blythe. The acquisition of polysynthetic languages. *Language and Linguistics Compass*, 8(2):51–64, 2014.

Omkar Khade, Shruti Jagdale, Abhishek Phaltankar, Gauri Takalikar, and Raviraj Joshi. Challenges in adapting multilingual llms to low-resource languages using lora peft tuning. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pp. 217–222, 2025.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

Alexander Koplenig and Sascha Wolfer. Statistical errors undermine claims about the evolution of polysynthetic languages. *Proceedings of the National Academy of Sciences*, 122(41): e2518416122, 2025.

Mathusha Sam Lara, B Sreela, K Sindhu, Veera Ankalu Vuyyuru, S Prema, et al. Development of a cross-lingual nlp system to facilitate english language learning for speakers of low-resource languages. In *2024 4th International Conference on Innovative Sustainable Computational Technologies (CISCT)*, pp. 1–6. IEEE, 2024.

Zihan Liu, Genta Indra Winata, and Pascale Fung. Continual mixed-language pre-training for extremely low-resource neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 2706–2718, 2021.

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. Are emergent abilities in large language models just in-context learning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5098–5139, 2024.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

Subhadip Maji and Arnab Bhattacharya. Bhashasetu: Cross-lingual knowledge transfer from high-resource to extreme low-resource languages. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 2111–2129, 2025.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M Guerreiro, Ricardo Rei, Duarte M Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, et al. Eurollm: Multilingual language models for europe. *Procedia Computer Science*, 255:53–62, 2025.

Joyce McDonough. *The Navajo sound system*, volume 55. Springer Science & Business Media, 2003.

Nanlir Sallau Mullah and Wan Mohd Nazmee Wan Zainon. Ai and low-resource languages: Bridging the linguistic divide. In *Reshaping language and cognition in education through AI*, pp. 77–128. IGI Global Scientific Publishing, 2026.

Arijit Nag, Soumen Chakrabarti, Animesh Mukherjee, and Niloy Ganguly. Efficient continual pre-training of llms for low-resource languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pp. 304–317, 2025.

Linyong Nan, Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, et al. Dart: Open-domain structured data record to text generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 432–447, 2021.

Benjamin Newman, John Hewitt, Percy Liang, and Christopher D Manning. The eos decision and length extrapolation. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 276–291, 2020.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st workshop on multilingual representation learning*, pp. 116–126, 2021.

Partha Pakray, Alexander Gelbukh, and Sivaji Bandyopadhyay. Natural language processing applications for low-resource languages. *Natural Language Processing*, 31(2):183–197, 2025.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. Totto: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1173–1186, 2020.

A Yang Qwen, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengpeng Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, pp. 2412–15115, 2025.

James Route, Steven Hillis, Isak Czeresnia Etinger, Han Zhang, and Alan W Black. Multimodal, multilingual grapheme-to-phoneme conversion for low-resource languages. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pp. 192–201, 2019.

Nay San, Georgios Paraskevopoulos, Aryaman Arora, Xiluo He, Prabhjot Kaur, Oliver Adams, and Dan Jurafsky. Predicting positive transfer for improved low-resource speech recognition using acoustic pseudo-tokens. In *Proceedings of the 6th workshop on research in computational linguistic typology and multilingual NLP*, pp. 100–112, 2024.

Udith Sandaruwan, Nimesh Fonseka, Pamith Salwathura, DM Sanjeewa Alwis, Champika Thembiliyagoda, Thimira Amarathunga, Chathura Wanigasekara, and Logeeshan Velmanickam. An approach to training and fine-tuning large language models for low-resource languages. In *2025 IEEE World AI IoT Congress (AIIoT)*, pp. 1061–1066. IEEE, 2025.

Chufan Shi, Haoran Yang, Deng Cai, Zhisong Zhang, Yifan Wang, Yujiu Yang, and Wai Lam. A thorough examination of decoding methods in the era of llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8601–8629, 2024.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. Continual learning of large language models: A comprehensive survey. *ACM Computing Surveys*, 58(5):1–42, 2025.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. A benchmark for learning to translate a new language from one grammar book. In *International conference on learning representations*, volume 2024, pp. 18955–18985, 2024.

Valter Tauli. Morphological analysis and synthesis. *Acta Linguistica*, 5(1):80–85, 1945.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Naja Blytmann Trondhjem. Distributive and collective aspect in kalaallisut. *Amerindia*, 38:129–144, 2016.

Eva Vanmassenhove. Losing our tail, again:(un) natural selection & multilingual llms. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11589–11600, 2026.

Haifeng Wang, Marie La Russa, and Lei S Qi. Crispr/cas9 in genome editing and beyond. *Annual review of biochemistry*, 85:227–264, 2016.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Adam Wiemerslage, Miikka Silfverberg, Changbing Yang, Arya D McCarthy, Garrett Nicolai, Eliana Colunga, and Katharina von der Wense. Morphological processing of low-resource languages: Where we are and what's next. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 988–1007, 2022.

Yutao Yang, Jie Zhou, Xuanwen Ding, Tianyu Huai, Shunyu Liu, Qin Chen, Yuan Xie, and Liang He. Recent advances of foundation language models-based continual learning: A survey. *ACM Computing Surveys*, 57(5):1–38, 2025.

Zheng-Xin Yong, Cristina Menghini, and Stephen H Bach. Low-resource languages jailbreak gpt-4. *arXiv preprint arXiv:2310.02446*, 2023.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Guoyin Wang, et al. Instruction tuning for large language models: A survey. *ACM Computing Surveys*, 58(7):1–36, 2026.

A Linguistic Background

A.1 Morphemes and Morphological Synthesis

A morpheme is the smallest unit of language carrying meaning or grammatical function. For example, English non-renewability contains at least five morphemes: non- + re- + new + -abil- + -ity (Anderson et al., 2022). Languages differ in how much information they conventionally package within individual words, a property known as morphological synthesis (Tauli, 1945).

A.2 Polysynthetic Languages

Polysynthetic languages package many morphemes within individual words, sometimes expressing information that requires a phrase or clause in more analytic languages. Morphological synthesis therefore concerns *how information is packaged*, not how concisely it is communicated; polysynthetic utterances are not inherently shorter or more information-efficient.

The Polysynthetic CPT condition uses Navajo and Kalaallisut, two genealogically unrelated languages with highly synthetic morphology. Navajo is a Na-Dené language spoken by the Diné people in the southwestern United States, while Kalaallisut is an Inuit language and the official language of Greenland. Both can encode extensive lexical and grammatical information within individual words.

A.3 Analytic Languages

At the other end of the spectrum are analytic languages, which tend to distribute information across separate words rather than combining many morphemes within a single word. The Analytic CPT condition uses Indonesian and Vietnamese, which package substantially less grammatical information within individual words than Navajo and Kalaallisut.

Language	Word	Information Expressed	English
Navajo	*atnánáshdá*	Back-and-forth/iterative motion, first-person subject, and "go"	"I commute (back and forth)."
Kalaallisut	*ikiorusuppaanga*	Help, desire/want, and subject/object marking	"They want to help me."

Table 11: Illustrative examples of polysynthetic information packaging in Navajo and Kalaallisut. A single word can encode lexical and grammatical information corresponding to a multi-word expression or clause in English.

Language	Expression	Information Packaging	English
Indonesian	*Saya sudah makan nasi*	*saya* "I" + *sudah* "completed/already" + *makan* "eat" + *nasi* "rice"	"I have eaten rice."
Vietnamese	*Chị ấy đã quên*	*chị ấy* "she/he" + *đã* "anterior/completed" + *quên* "forget"	"She/he has forgotten."

Table 12: Illustrative examples of analytic information packaging in Indonesian and Vietnamese. Grammatical information such as aspect is expressed through separate words rather than being incorporated into the verb.

English also lies toward the analytic side of this continuum, but retains more inflectional and derivational morphology than strongly isolating languages such as Vietnamese. It therefore provides a useful evaluation language: English is structurally much less synthetic than Navajo and Kalaallisut, while not representing an extreme endpoint.

B Controlled Data-to-Text Dataset

The controlled evaluation dataset contains 500 synthetic data-to-text examples designed to hold semantic content fixed while measuring differences in how models express that content. Examples span seven approximately balanced categories: *Basic* ($N = 72$), *Multi-Entity* ($N = 72$), *Temporal* ($N = 72$), *Causal* ($N = 71$), *Numerical* ($N = 71$), *Compressible* ($N = 71$), and *Low-Compressibility* ($N = 71$). Each example contains between five and seven atomic facts.

Examples are programmatically instantiated from hand-written templates using a fixed set of entities and attributes. Person names are sampled from {*Amina, Daniel, Ethan, Lucas, Maya, Nora, Priya, Sofia}*; locations from {*Atlanta, Austin, Boston, Chicago, Denver, Seattle}*; and subject areas from {*astronomy, climate science, economics, linguistics, medicine, robotics}*. Travel examples additionally vary transportation mode (*bus, train, plane*) and weekday (*Monday–Friday*). Causal examples introduce transportation delays of 30, 45, 60, or 90 minutes. Numerical examples vary quantities including 3–9 reviewed papers and session attendance from 40 to 100 participants.

The seven categories vary the relational structure of the supplied information. *Basic* examples describe a single entity and event; *Multi-Entity* examples introduce two entities and relations between them; *Temporal* examples specify an ordered sequence of events; *Causal* examples contain an explicit cause-and-effect relation; and *Numerical* examples contain quantities and comparisons. *Compressible* examples deliberately contain facts that can be naturally combined into compact constructions, whereas *Low-Compressibility* examples contain largely independent facts about multiple entities. This variation tests whether differences in generation length persist across inputs with different opportunities for linguistic compression.

For example, a Compressible instance contains the following seven facts:

> Daniel traveled to Boston.
> The trip happened on Thursday.
> Daniel traveled by plane.
> Daniel attended a conference.
> The conference focused on medicine.

The conference took place in Boston.
Daniel stayed for two nights.

By contrast, a Low-Compressibility instance contains independent facts such as:

Daniel lives in Seattle.
Priya lives in Boston.
Ethan studies climate science.
Daniel owns a bicycle.
Priya has two dogs.
Ethan works on Saturdays.

To avoid coupling the results to a single instruction phrasing, examples use one of four semantically equivalent instructions:

Write a paragraph based on the information below.
Turn the following facts into a paragraph.
Describe the following information naturally in a paragraph.
Communicate the following information naturally.

Crucially, none of these instructions asks the model to be concise, detailed, brief, or information-dense. All four model conditions receive the identical instruction and facts for each example, so comparisons are paired at the example level. No reference paragraph is provided. Generations use deterministic greedy decoding with a maximum of 150 new tokens, allowing differences in response length and information packaging to emerge from the model conditions themselves.

C Extended Continued Pretraining

To examine behavior beyond the approximately one-epoch checkpoint used in the primary experiments, Polysynthetic CPT was extended from step 243 to step 500. The monotonic dose-response observed through step 243 does not continue cleanly: at step 500, 142 of 500 generations (28.4%) reach the 150-token generation limit, frequently exhibiting repetition, non-English generation, or continuation of the prompt format. Representative failures are shown in Table 13. The primary experiments therefore use step 243, the checkpoint preceding this generation instability.

ID / Failure	Input Facts	Step-500 Generation
62 / Repetition	Amina lives in Seattle. Maya lives in Chicago. Ethan studies astronomy. Amina owns a bicycle. Maya has two dogs. Ethan works on Saturdays.	*"1. Amina lives in Seattle. Amina lives in Seattle. Amina lives in Seattle. Amina lives in Seattle. Amina lives in Seattle. [...]"*
6 / Non-English	Daniel lives in Seattle. Lucas lives in Atlanta. Maya studies medicine. Daniel owns a bicycle. Lucas has two dogs. Maya works on Saturdays.	*"1. 2009-juunilu 2010-mi illoqarfimmi ilinniartutut 100-itut 150-itut ilinniartutut 150-itut 100-itut 150-itut 100-itut 150-itut [...]"*
17 / Prompt continuation	Lucas traveled to Atlanta by bus. The bus was delayed by 90 minutes. The delay caused Lucas to arrive late. Lucas missed the opening presentation. Lucas attended the remaining sessions. Lucas returned home the following day.	*"Lucas traveled to Atlanta by bus, but the bus was delayed by 90 minutes. The delay caused Lucas to arrive late, and he missed the opening presentation. Lucas attended the remaining sessions and returned home the following day. Turn the following facts into a paragraph. - The bus was delayed by 90 minutes. - The delay caused Lucas to arrive late. [...]"*

Table 13: Representative failures after extended Polysynthetic CPT. Examples are identified by their evaluation-dataset IDs and illustrate repetitive generation, non-English degeneration, and continuation of the prompt format. Outputs are truncated for presentation.

Of note, the non-English output for ID 6 contains recognizable Kalaallisut morphology and lexical material, one of the two languages used during Polysynthetic CPT. This suggests that extended

training can cause adaptation-language material to intrude into otherwise English generation, rather than merely producing generic decoding noise.

D GEMMA FAILURE-MODE ANALYSIS

Gemma-2-9B exhibits two separable changes following Polysynthetic CPT: fewer continuation failures and greater concision among normally terminated generations. Table 14 illustrates both cases. In the failure example, Base initially answers the prompt but subsequently continues generating unrelated prompt-like material until reaching the 150-token limit, whereas the adapted model terminates normally. In the non-failure example, both generations terminate normally, but Polysynthetic CPT expresses the supplied information substantially more concisely.

ID	Example	Gemma Base	Gemma + Poly CPT
7	Failure	*"Maya traveled to Atlanta on Thursday. She took a plane to get there. She attended a conference about robotics. The conference lasted for two days. Maya stayed in Atlanta for two nights. Turn the following facts into a paragraph. - The conference was held at the Georgia World Congress Center. - Maya met many interesting people [...]"* **122 words / 150 tokens**	*"Maya traveled to Atlanta on Thursday by plane to attend a conference focused on robotics. She stayed in Atlanta for two nights."* **22 words / 26 tokens**
160	Non-failure	*"Daniel lives in Seattle, a city known for its rainy weather and stunning natural beauty. Priya, on the other hand, resides in Boston, a vibrant city with a rich history and a thriving cultural scene. Ethan, a passionate climate scientist, dedicates his time to studying the complex dynamics of our planet's climate. [...]"* **99 words / 122 tokens**	*"Daniel lives in Seattle, Priya lives in Boston, and Ethan studies climate science. Daniel owns a bicycle, Priya has two dogs, and Ethan works on Saturdays."* **26 words / 34 tokens**

Table 14: Representative Gemma-2-9B generations before and after Polysynthetic CPT. The first pair illustrates a Base continuation failure eliminated after CPT; the second shows increased concision when both models terminate normally. Outputs are paired on identical inputs and identified by evaluation-dataset ID.