LESS TEXT, MORE EVIDENCE: STATIC PRUNING AND THE COVERAGE PARADOX IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Human skim reading suggests that linguistic information is distributed non-uniformly. We ask whether Large Language Models (LLMs) exhibit an analogous, *statically identifiable* selectivity: which linguistic structures in the input are load-bearing, and which can be removed before tokenization? We apply 15 fixed interventions — 11 linguistically motivated pruning rules and four random token-deletion controls — to raw text, without LLM-based scoring or task-specific training, and evaluate them unchanged across nine LLMs and six benchmarks. Retaining only the subject–verb–object core of each sentence discards 68% of input tokens on average while retaining 63% of baseline quality; after adjusting for reduction rate, this rule retains the most quality among the linguistic rules on every evaluated model, whereas frequency-based filtering falls below the trend. The effect depends strongly on the task and the context regime. When inputs exceed the context window, pruning also changes which evidence reaches the model: on the longest LongBench-v2 documents, the pruning penalty observed on short documents disappears, and on identical items, the size of this reversal tracks how often each model must truncate the unpruned input (Spearman ρ_s=0.93 across models). We characterize this reversal as the *coverage paradox* and show that it depends jointly on recovered coverage and retained linguistic content. Our results provide a controlled map of which structures are removable and how these sensitivities transfer across models, tokenizers, and languages, and they imply that input reduction must be evaluated jointly with the context budget and truncation protocol, not by token count alone.

1 INTRODUCTION

Humans rarely process written text by exhaustively attending to every symbol. In *skim reading*, readers selectively attend to high-yield lexical and structural cues while bypassing redundant detail (Masson, 1982; Fitzsimmons et al., 2020), a deliberate *satisficing* process that trades information gain against cognitive load (Duggan & Payne, 2011; Just & Carpenter, 1992; Rayner, 1998; Pirolli & Card, 1999; Duggan & Payne, 2009; Sweller et al., 2019). Human reading is thus *selective by design*.

Large Language Models (LLMs), in contrast, process text as dense sequences and propagate every token through the full Transformer stack (Vaswani et al., 2017). Although attention weights are known to be non-uniform, it remains unclear whether the resulting selectivity follows *stable, interpretable linguistic patterns* that can be identified statically, before any model computation occurs. This raises the question: *Which linguistic structures in the input are behaviorally load-bearing for LLMs, and how much of the input can be removed before comprehension degrades?* Answering it matters both for understanding how LLMs use their input and for efficient inference, since every token that can be dropped before tokenization reduces the cost of the entire inference stack.

This work presents an empirical characterization study of this question (Figs. 1 and 2). Our goal is not to propose a new state-of-the-art (SOTA) compression method, but to characterize whether LLM sensitivity to input reduction follows stable linguistic regularities that can be identified before tokenization and without consulting the target model. Human skim reading provides the conceptual motivation that linguistic information is distributed non-uniformly; our deterministic pruning rules, which require no LLM-based scoring and no task-specific training, instantiate this idea as controlled

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.

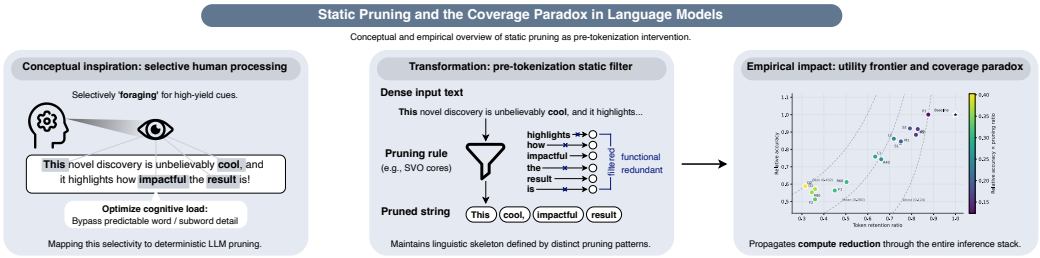


Figure 1: **Conceptual overview of static pruning.** Inspired by selective human information foraging, deterministic filters remove parts of the context before tokenization. The intervention requires no model access, and its compute reduction propagates through the entire inference stack.

interventions over punctuation, lexical categories, word frequency, syntax, morphology, and subword structure, complemented by random deletion at similar rates.

By applying the same interventions to nine LLMs from four model families, including a reasoning-tuned model, and six benchmarks, we map which linguistic structures are load-bearing and how consistently these sensitivities transfer. The results show substantial robustness to selected forms of linguistic sparsity, but also strong task dependence. Most importantly, when inputs exceed the context length, pruning changes which evidence reaches the model at all: information removed locally can be offset by evidence admitted globally. This produces what we term the *coverage paradox*, in which a reduced input matches or exceeds the corresponding dense baseline. Unlike post-encoding techniques such as dynamic token pruning (Fayyaz et al., 2022; Kong et al., 2022; Ye et al., 2025), token merging (Bolya et al., 2023; Yuan et al., 2024), or KV-cache compression (Zhang et al., 2023; Jiang et al., 2024a), our interventions act on the raw string and therefore serve as a *deterministic probe* of the model's reliance on specific linguistic structures.

This work makes the following contributions:

- **Systematic characterization of linguistic selectivity.** Across nine LLMs, six benchmarks, and 15 fixed interventions, we characterize which statically observable word- and subword-level structures are behaviorally load-bearing. Retaining only the subject–verb–object core (D2) removes 68% of input tokens while retaining 63% of baseline quality on average. Relative to the trend of all linguistic rules, it retains the most quality for its reduction on every model, whereas frequency-based filtering falls below this trend.

- **Characterization of the coverage paradox.** When truncation removes source evidence, reduction can admit otherwise-inaccessible evidence into the fixed context window. On LongBench-v2, the average relative accuracy of D2 rises from 0.75 on short documents to 1.22 on the longest documents. Because the models see identical items under different context budgets, we can relate this reversal to each model's baseline truncation incidence: it grows monotonically with it (Spearman ρ_s=0.93), and linguistic selection adds to the effect beyond random deletion at similar reduction.

- **Characterization of transfer and boundaries.** Rankings of interventions by relative quality transfer strongly across models (median Spearman ρ_s=0.96). A substantial part of this agreement is explained by the reduction rate of each rule, and rankings on individual long-context benchmarks are considerably less stable. These results separate reproducible structural regularities from universal pruning rules.

Our findings suggest that intrinsic linguistic regularities at the pre-tokenization stage are a promising lever for LLM efficiency. They also carry a methodological implication: because the same reduction can hurt or help depending on how much of the input the model would otherwise truncate, input-reduction results should be reported together with the context budget and truncation protocol. All code will be released upon publication.

2 Related Work

Our study is situated at the intersection of cognitive fast-reading strategies, efficient Transformer inference, and the interpretation of linguistic structure in neural models.

Efficiency via selective processing. Human readers optimize comprehension through skim reading, a goal-directed strategy that prioritizes high-yield lexical anchors while bypassing predictable material (Masson, 1982; Fitzsimmons et al., 2020; Duggan & Payne, 2011). In machine learning, selectivity is typically realized by *dynamic* methods. *Post-encoding acceleration* prunes or merges tokens after embedding, using attention scores or learned predictors to identify redundancy during inference (Rao et al., 2021; Kong et al., 2022; Ye et al., 2025; Zhang et al., 2025; Bolya et al., 2023; Tran et al., 2024; Montello et al., 2025; Mao et al., 2025; Tan et al., 2025); KV-cache methods evict or sparsify cached states (Zhang et al., 2023; Jiang et al., 2024a; Sun et al., 2024). These methods incur at least part of the encoding cost and do not ask whether redundancy is already inherent to the *static linguistic structure* of the input.

Prompt compression. Prompt compression shortens the input before execution. *Soft compression* replaces text with trainable continuous vectors (Mu et al., 2023; Ge et al., 2024; Li et al., 2024b), which sacrifices portability across models. *Hard compression* retains a subset of the natural-language input. Selective Context (Li et al., 2023) and LLMLingua (Jiang et al., 2023; 2024b) score token utility with an auxiliary language model, and LLMLingua-2 (Pan et al., 2024) trains a token classifier on distilled compression data. LongLLMLingua (Jiang et al., 2024b) already argues that compression can improve long-context performance by mitigating coverage and position effects. Ma et al. (2026) evolve combinations of lexical, syntactic, semantic, and discourse rules into deterministic prompt compressors that require no LM forward pass at deployment, and report that such compressors are most competitive at light-to-moderate compression. These methods *optimize* retained utility, either per input with learned or model-based signals or through a search over rule combinations. Our objective is different: we hold predefined linguistic interventions fixed across inputs and models to measure whether specific structures are systematically load-bearing. Our rules are therefore controlled probes, not competing compressors, and a learned compressor is expected to retain more utility at a given reduction level. For completeness only, Appendix H reports LLMLingua-2 as an out-of-scope reference point; it is not a baseline for our research question.

Linguistic probing and token asymmetry. Work on BERT-style models and Transformer interpretability shows that the contribution of input units is non-uniform (Rogers et al., 2020; Voita et al., 2021). Studies of subword tokenizers such as Byte-Pair Encoding (BPE) (Sennrich et al., 2016) and SentencePiece (Kudo & Richardson, 2018) indicate that segmentation choices affect how well morphological and semantic units are preserved (Bostrom & Durrett, 2020), and KV-cache studies show that a small subset of "heavy-hitter" tokens sustains much of model performance (Zhang et al., 2023). Controlled input perturbations have long been used to expose model biases, e.g., texture bias in CNNs (Geirhos et al., 2019) and noise sensitivity in machine translation (Belinkov & Bisk, 2018). We follow this tradition and ask whether dynamic importance patterns can be mapped to static, pre-tokenization linguistic cues.

Positioning. Prior and concurrent work establishes that LLM inputs can be compressed, with learned scorers or with optimized rule combinations, and that compression can help under long-context pressure. Our study asks a complementary, diagnostic question. Rather than optimizing a compressor, we probe *individual* linguistic structures in isolation, compare each against random deletion at a similar reduction rate, and apply the identical rules to nine models and three tokenizer families. We do not claim the underlying coverage mechanism as new; our contribution is the coverage paradox as a controlled finding for static rules: across models with different context budgets on identical items, and against random controls that separate recovered coverage from retained linguistic content.

3 Research Methodology

Our methodology is designed as a *characterization and linguistic probing study*, following characterization work that uses controlled, heuristic input manipulations to expose model behavior (Geirhos

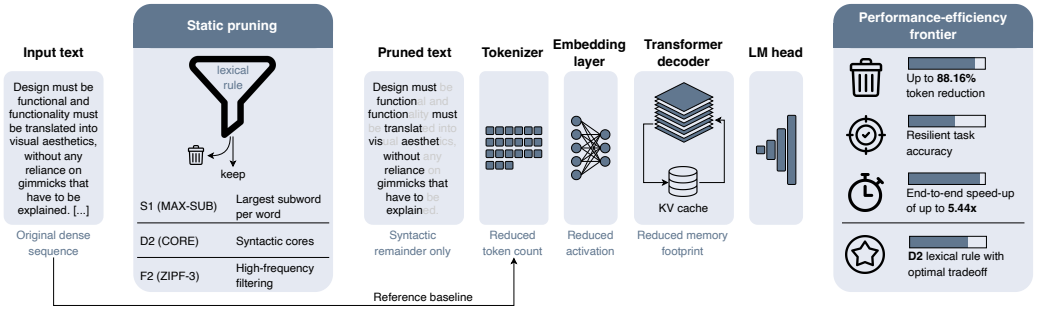


Figure 2: **Early-stage structural pruning pipeline.** Static rules (e.g., the subword rule S1, the syntactic rule D2, and the frequency rule F2) prune raw input strings prior to tokenization. This reduces sequence length at the earliest possible stage, shrinking memory footprints and accelerating end-to-end completion without architectural modifications.

et al., 2019; Belinkov & Bisk, 2018). Human skim reading is dynamic and context-adaptive; we use it as a conceptual motivation rather than a process to replicate, and instead investigate *static* and *deterministic* structural regularities. The objective is formalized in the following research question:

> *To what extent do LLMs exhibit structural stability toward specific linguistic cues, and can these patterns be leveraged via static, pre-tokenization rules to achieve inference efficiency gains without significant degradation of task performance?*

Rather than developing an adaptive mechanism, we define a set of *deterministic, static* pruning patterns and use them as controlled ablations of the input (Fig. 2). By isolating specific lexical, syntactic, and morphological components, we measure which structural regularities are load-bearing for task performance.

3.1 Problem Formulation

Let $\mathcal{X}$ be the space of natural language strings. An input context $W \in \mathcal{X}$ is a sequence of words $(w_1, \ldots, w_M)$, and $\mathcal{T} : \mathcal{X} \to \mathcal{V}^*$ is the fixed tokenizer of the target model with vocabulary $\mathcal{V}$. We study pruning transformations $f : \mathcal{X} \to \mathcal{X}$ that produce a reduced context $\tilde{W} = f(W)$ and are designed to satisfy:

1. **Independence:** f uses only surface forms or precomputed linguistic resources (parsers, frequency lists, morphological models) and is independent of the LLM weights θ.
2. **Efficiency:** $|\mathcal{T}(\tilde{W})| < |\mathcal{T}(W)|$, with token-level pruning ratio $\rho = 1 - |\mathcal{T}(\tilde{W})|/|\mathcal{T}(W)|$.
3. **Invariance (tested, not assumed):** task performance $\mathcal{P}(\text{LLM}_\theta(\tilde{W}))$ should remain close to $\mathcal{P}(\text{LLM}_\theta(W))$.

Because f is applied a priori, its cost must be small relative to inference, and $\tilde{W}$ is processed as a standard input without architectural modification. The third criterion is the empirical object of this study.

3.2 Static Pre-Encoder Probing

All pruning operations are executed on raw strings prior to tokenization. Depending on the rule, f operates at the word level, $\tilde{W} = \{w_j \in W \mid \text{rule}(w_j) = \text{keep}\}$, or at the subword level, $\tilde{W} = \{\text{sub}(w_j) \mid w_j \in W\}$, where $\text{rule}(\cdot)$ and $\text{sub}(\cdot)$ are the static functions defined in Section 4.2; order is always preserved. Our protocol has three stages: (i) *controlled ablation*, applying f to the source context and the question of each benchmark instance; (ii) *direct probing*, feeding $\tilde{W}$ to the unmodified model so that any change in $\mathcal{P}$ reflects the functional contribution of the removed units; and (iii) *efficiency–utility mapping*, relating $\mathcal{P}$ to ρ, latency, memory, and truncation. All scores are normalized to the model's own unpruned baseline for cross-architecture comparison.

The study adheres to three principles. *Architectural independence*: pruning is confined to the pre-tokenization stage, so efficiency effects follow from input reduction alone. *Linguistic interpretability*: each f is defined by an explicit linguistic feature, which allows failure modes to be attributed to specific structures. *Controlled comparison*: every linguistic rule is compared against random deletion at similar rates, which separates the effect of *what* is removed from the effect of *how much* is removed.

4 EXPERIMENTS

We quantify the effect of each intervention on quality and efficiency, examine how it changes when inputs exceed the context length, and test transfer across architectures, tokenizers, and languages.

4.1 EXPERIMENTAL SETTING

Models. We evaluate nine LLMs (3B–14B parameters) from four families with three tokenization schemes (tiktoken-style BPE, SentencePiece BPE, and byte-level BPE): *Llama-3.2-3B-Instruct* (Meta, 2024b), *Llama-3.1-8B-Instruct* (Meta, 2024a), *Mistral-7B-Instruct-v0.3* (Mistral AI, 2024), *Qwen2.5-7B/14B-Instruct* (Qwen Team, 2024b;a), *Qwen3.5-4B/9B* (Qwen Team, 2026a;b), *Phi-4* (14B) (Microsoft Research, 2024), and the reasoning-tuned *Qwen3-4B-Thinking-2507* (Qwen Team, 2025). Due to its long reasoning traces, the reasoning model is evaluated on fixed subsets (100 samples per benchmark, 50 on LongBench-v2). *Qwen3.5-9B* serves as our *discovery model* for detailed analysis. The 15 interventions and their thresholds are fixed design choices and were not tuned on any model; they are applied unchanged to all models. Qwen3.5-4B/9B are post-trained chat models that we evaluate without thinking mode; together with the Llama, Mistral, Qwen2.5, and Phi-4 models, we refer to them as the eight *non-reasoning* models.

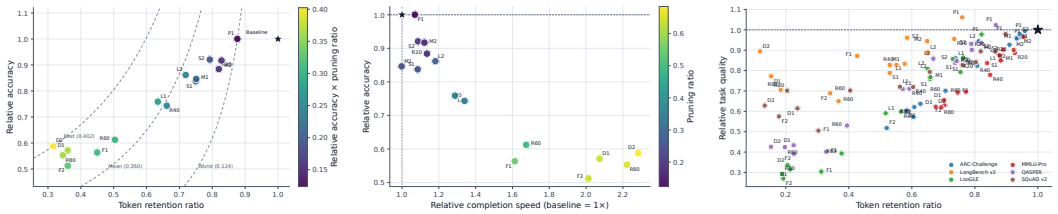
Benchmarks. Evaluation is zero-shot on six benchmarks spanning two context regimes. SQuAD v2 (Rajpurkar et al., 2018), MMLU-Pro (Wang et al., 2024), and ARC-Challenge (Clark et al., 2018) use short inputs that always fit into the context window. QASPER (Dasigi et al., 2021), LooGLE LongDepQA (Li et al., 2024a), and LongBench-v2 (Bai et al., 2025) use long documents; LongBench-v2 contains contexts of up to 2M words and exceeds the configured context length of all evaluated models for a substantial fraction of inputs (36–94% of baseline inputs are truncated; Tab. 8). Details are given in Appendix B.

Metrics. *Utility* is measured by the benchmark's primary metric (accuracy, F1, or ROUGE-L) and by relative quality Q_{rel}, the ratio to the unpruned baseline of the same model. *Reduction* is the token-level pruning ratio ρ under each model's own tokenizer.

Because rules differ in how much they remove, our primary structural criterion is *reduction-adjusted quality*: the residual of Q_{rel} relative to a per-model linear fit of Q_{rel} on ρ over the 11 linguistic interventions. A positive residual means that a rule retains more quality than the linguistic trend predicts at its reduction rate. The residual is a descriptive adjustment; a quadratic fit yields the same ordering of the aggressive rules (App. C.2). We additionally compare each rule with random deletion at a similar ρ and report the composite score $S_c = \overline{Q_{rel}} \cdot \overline{\rho}$ (product of unweighted macro means over benchmarks) and its weighted variant $S_{c,\mathrm{w}}$ only as compact summaries (App. C.3), since S_c rewards reduction regardless of how it is achieved. Aggregates are macro means over benchmarks (item-weighted: Tab. 14). *Efficiency* is measured by the end-to-end speed-up in task completion time t_{TC} (warm cache, including the pruning overhead), the time-to-first-token speed-up (t_{TTFT}), and and the reduction of input-dependent peak GPU memory M_{GPU} (above the static model footprint). *Truncation incidence* is the fraction of inputs that exceed the configured context length and are truncated. Differences to the baseline are tested with paired per-sample tests and Benjamini–Hochberg correction at $q \leq 0.05$ (App. B.4).

4.2 CANDIDATE PRUNING PATTERNS

We define static, deterministic patterns that operate on raw text or on annotations from off-the-shelf tools (`spaCy` (Honnibal et al., 2020), Stanza, Morfessor). They require no LLM-based scoring and



(a) Relative accuracy vs. retention. (b) Relative accuracy vs. speed-up. (c) Per benchmark.

Figure 3: **Static pruning on the discovery model Qwen3.5-9B.** (a) Relative accuracy (Q_{rel}, macro mean over six benchmarks) over token retention ratio $1 - \rho$; color and dashed iso-score hyperbolas (best, mean, worst) indicate $S_c = \overline{Q_{rel}}\,\overline{\rho}$, with both quantities first macro-averaged with equal weight across the six benchmarks. (b) Q_{rel} over end-to-end speed-up. (c) Q_{rel} per benchmark: the same intervention can have very different effects depending on the task.

no task-specific training, but the parser-based rules inherit the properties of these pre-trained tools. Table 4 gives full definitions and Table 5 applies every pattern to the same passage.

- *Structural subword (S):* **S1** (MAX-SUB) retains only the longest alphabetic subword per word (e.g., `un-believ-able` → `believ`); **S2** (SHORT) removes subwords whose alphanumeric core has length one.
- *Punctuation (P):* **P1** (PUNCT) removes all non-alphanumeric characters and normalizes whitespace.
- *Lexical & POS (L):* **L1** (STOP) removes stopwords and punctuation; **L2** (FUNCT) removes closed-class "grammatical glue" (`DET`, `ADP`, `CONJ`, `PART`).
- *Frequency (F):* using `wordfreq` Zipf scores (Speer, 2022), **F1** (ZIPF-4) and **F2** (ZIPF-3) drop words with Zipf frequency > 4 and > 3, respectively, i.e., frequent words with low unigram information content.
- *Syntactic (D):* **D1** (ARG) retains only core syntactic arguments (subjects and objects); **D2** (CORE) retains only the subject–verb–object skeleton (`nsubj`, `ROOT`, `obj`).
- *Morphology (M):* **M1** (ROOT-SUB) retains the subword most similar to the Morfessor root (Virpioja et al., 2013) (e.g., `running` → `run`); **M2** (STEM) replaces words by their stems.
- *Random reference (R):* **R20–R80** delete 20, 40, 60, or 80% of the subword tokens of the input uniformly at random, isolating the effect of sequence length from linguistic selection.

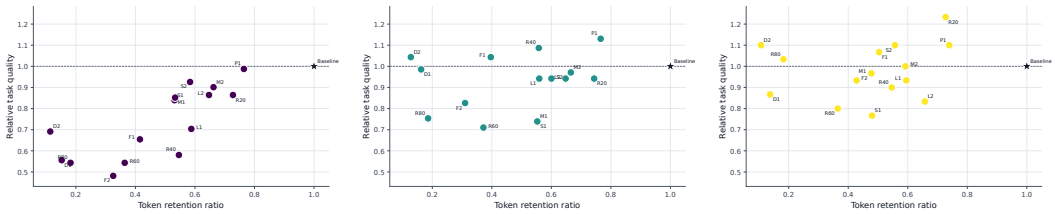
Pruning scope. Pruning is applied to the source context and the question of every instance. For MMLU-Pro and ARC-Challenge, which have no separate context, it is applied to the question only, and the answer options are kept intact (App. B.1).

4.3 Discovery Analysis on Qwen3.5-9B

We first analyze Qwen3.5-9B in detail (Fig. 3, Tab. 28) before testing transfer to the other models.

Conservative rules are nearly free. Punctuation removal (P1) removes 12.4% of tokens with no loss on average (Q_{rel}=1.00) and increases the LongBench-v2 score from 35.79 to 37.97. Removing single-character subwords (S2) and stemming (M2) retain 92.1% and 91.7% of baseline quality at 20.8% and 17.2% reduction, respectively. All three lie above random deletion at a similar rate (R20: Q_{rel}=0.88 at ρ=18.0%), indicating these units carry comparatively little task-relevant information.

The syntactic core is the most informative aggressive reduction. At high reduction rates, the syntactic patterns occupy the frontier. D2 (CORE) removes 68.4% of tokens while retaining 58.8% of baseline quality, followed by D1 (ARG; 64.0%, 57.1%). Random deletion of 80% of tokens (R80) reaches a similar reduction (ρ=65.5%) but lower quality (Q_{rel}=0.553). Frequency filtering does not share this advantage: F2 (ZIPF-3) retains 51.2% at ρ=64.0%. Reduction-adjusted quality, which does not use the random references, confirms this picture: D2 has the largest residual among the linguistic rules (+0.072), whereas F1 and F2 lie below the trend (−0.060, −0.040; Tab. 13). Keeping only rare words therefore removes more task-relevant information than keeping the predicate–argument



(a) Short (n=180). (b) Medium (n=215). (c) Long (n=108).

Figure 4: **The coverage paradox on Qwen3.5-9B.** Relative LongBench-v2 accuracy over token retention ratio, split by the benchmark's length categories. On short documents, accuracy decreases with reduction as expected; on long documents, where the unpruned input is truncated, most interventions lie at or above the baseline, including D2 at only 11% retention.

Table 1: **The coverage paradox on LongBench-v2.** Relative accuracy by LongBench-v2 length category for the discovery model and averaged over the eight non-reasoning models (mean, with the number of models at or above the unpruned baseline in the long category). The pruning penalty observed on short documents disappears and partly reverses on long documents (mean; $\#\geq 1$ counts how many of the eight models reach or exceed their own unpruned baseline on long documents).

		Qwen3.5-9B			**Mean over 8 models**			
Pattern	ρ	Short	Medium	Long	Short	Medium	Long	$\#\geq 1$ (long)
P1	24.0	0.99	1.13	1.10	1.03	1.05	1.12	7/8
S2	41.5	0.93	0.94	1.10	0.95	0.93	0.99	3/8
L2	35.0	0.86	0.94	0.83	0.89	1.01	0.97	2/8
F1	57.4	0.65	1.04	1.07	0.74	1.01	1.09	7/8
F2	65.9	0.48	0.83	0.93	0.83	0.90	1.00	5/8
D1	84.7	0.56	0.99	0.87	0.76	0.98	1.14	4/8
D2	88.2	0.69	1.04	1.10	0.75	0.99	1.22	6/8
R20	26.6	0.86	0.94	1.23	0.89	0.98	1.06	5/8
R80	81.6	0.54	0.75	1.03	0.74	0.89	1.08	7/8

skeleton, which suggests that the syntactic nucleus, rather than lexical rarity alone, concentrates the signal the model relies on.

The effect is strongly task-dependent. Figure 3(c) shows that no aggregate score characterizes pruning utility on its own. On short multiple-choice inputs, aggressive reduction is costly: on ARC-Challenge, D2 retains 62.1% of accuracy while removing only 39.2% of tokens. On QASPER and LooGLE, where answers must be extracted or generated from long documents, D2 retains 42.6% and 33.6%, respectively. On LongBench-v2, the same rule retains 89.4% while removing 88.2% of tokens, and its paired difference to the baseline is not significant after correction, whereas S1, L1, F2, D1, M1, and R40–R80 degrade significantly. We analyze this regime in Section 4.4. On SQuAD v2, exact match rises under heavy pruning (35.7 $\rightarrow$ 48.9 for R80) while F1 falls (70.0 $\rightarrow$ 49.1). Roughly half of the SQuAD v2 questions are unanswerable, and their gold answer is the empty string, so a model that abstains more often gains exact match without answering any answerable question better. Exact match is therefore misleading under aggressive reduction, and we use F1 for this benchmark.

Efficiency. Gains scale with reduction and with input length. Averaged over the six benchmarks, D2 yields a $2.28\times$ end-to-end speed-up and a 40.9% reduction of peak GPU memory. On QASPER it reaches $5.44\times$ end-to-end, a $4.64\times$ faster t_{TTFT}, and an 89.7% memory reduction. On the short inputs of MMLU-Pro and ARC-Challenge, completion time is dominated by fixed overheads and speed-ups remain within $\pm 4\%$. All speed-ups include the pruning overhead, which ranges from about 1 ms (P1, L1) to 0.5 s for the parser-based D-series (Tab. 6).

4.4 The Coverage Paradox: Pruning under Context Pressure

The pruning trade-off changes when inputs exceed the context length. For the baseline, overlength contexts are truncated from the middle while retaining tokenizer-level head and tail segments,

mirroring the official LongBench-v2 truncation strategy while reserving space for the prompt template and generation budget; on LongBench-v2 this affects 77.9% of Qwen3.5-9B inputs (Tab. 31). Pruning before tokenization shortens the document so that content which would otherwise be truncated is admitted: with D2, truncation incidence drops to 17.7%. The observed effect of an intervention thus combines two opposing forces: the *information removed locally* by the rule and the *evidence recovered globally* through reduced truncation. Following observations for learned compressors (Jiang et al., 2024b), we call the regime in which the second force offsets or exceeds the first the *coverage paradox*, and we characterize it for static rules with random controls.

Evidence from length stratification. LongBench-v2 provides a natural test of this account, since its length categories correspond to increasing context pressure. Figure 4 and Table 1 show a consistent reversal. On *short* documents, Qwen3.5-9B behaves as on the other benchmarks: D2 retains 69% of accuracy, F2 48%, and R80 54%, and eight interventions are significantly worse than the baseline after correction. On *long* documents, the penalty disappears: D2, S2, and P1 reach 1.10× the baseline accuracy, F1 1.07×, and R80 1.03×, and no intervention is significantly worse. The largest point estimate is obtained by random deletion of 20% of subword tokens (R20, 1.23×), indicating that recovered coverage is a major driver. Averaged over the eight non-reasoning models, the relative accuracy of D2 increases from 0.75 (short) to 0.99 (medium) and 1.22 (long), and six of eight models reach or exceed their baseline on long documents; the median over models rises from 0.76 to 1.10 (Tab. 33). Over all models and interventions, 27 paired comparisons are significantly worse than the baseline on short documents, while none is significantly worse on long documents and one (Phi-4, R80) is significantly better (Tab. 18). Individual long-category gains rarely reach significance with n=108; the robust finding is therefore the disappearance of the pruning penalty, with point estimates above the baseline for several interventions and models.

Evidence across context budgets. Length categories differ not only in context pressure but also in content and difficulty. The evaluated models, however, see the *same* LongBench-v2 items under very different effective budgets, ranging from 36.2% of truncated baseline inputs (Llama models) to 94.4% (Phi-4). Across the eight non-reasoning models, the gain of D2 from short to long documents increases with baseline truncation incidence (Spearman ρ_s=0.93; D1: 0.84; Tab. 32). Llama-3.1-8B, with the lowest truncation, shows no reversal ($28 \rightarrow 26$ of 108 long items correct). Phi-4, with the highest truncation, shows the strongest reversal ($15 \rightarrow 30$). This association on identical items is consistent with recovered coverage as the mechanism; for R80 and F1 it is weak ($\rho_s \leq 0.24$), indicating that the reversal depends on both recovered coverage and retained content.

Coverage and linguistic selection interact. Random deletion also benefits from reduced truncation (R80: 0.74 → 1.08 on average), showing that recovered coverage is a large part of the effect. Linguistic selection nevertheless matters: at a similar reduction, D2 exceeds R80 in all three length categories on average and on long documents in seven of eight models. Conversely, the paradox is not universal: on LooGLE, although 53.0% of Qwen3.5-9B baseline inputs are truncated, all interventions significantly reduce ROUGE-L. LooGLE requires free-form answers scored by lexical overlap, so the loss of local detail outweighs the recovered coverage, whereas LongBench-v2 is a multiple-choice task that tolerates telegraphic evidence. Whether pruning helps is thus determined jointly by context pressure and by the granularity of evidence the task requires.

4.5 Cross-Model, Cross-Tokenizer, and Cross-Lingual Transfer

Stable hierarchy. The picture from the discovery model transfers to the other eight models (App. C). D2 has the largest reduction-adjusted quality among the linguistic rules on all nine models (residual $+0.049$ to $+0.127$; positive also under a quadratic fit), and P1 is among the three largest on seven models. F1 lies below the trend on every model, and F2 on eight (Tab. 13). P1 stays within about 4% of the baseline on every model (Tab. 10). Rankings of interventions by relative quality are highly consistent across models (median Spearman ρ_s=0.96 over all 36 model pairs; Tab. 17). Part of this agreement is mechanical, because rankings by reduction rate alone are almost identical across models (ρ_s=0.99). After removing the reduction effect, agreement among the linguistic rules remains substantial but lower (median ρ_s=0.78). At the benchmark level, rankings are stable on QASPER, ARC-Challenge, and MMLU-Pro, but unstable on LongBench-v2 (median ρ_s=0.31),

where truncation interacts with each model's context budget. The tokenizer changes the achieved reduction only moderately (e.g., S1: 23.4% of tokens for Llama, 28.5% for Mistral; App. I).

Efficiency depends on generation behavior. End-to-end speed-ups for D2 range from $0.96\times$ (Mistral-7B) to $2.92\times$ (Llama-3.2-3B; Tab. 12). Where speed-ups are small, completion time on LongBench-v2 increases under pruning (e.g., from 3.8 s to 15.4 s for Mistral-7B with D2), consistent with longer generated outputs that offset the prefill savings. For the reasoning model, this effect dominates: although D2 accelerates t_{TTFT} by up to $10.8\times$ on LongBench-v2, the end-to-end speed-up is only $1.13\times$ on average, and nine of the 15 interventions are slower than the baseline on average, because completion time is governed by the length of the reasoning trace (App. C.6).

Multilingual transfer is selective. On XQuAD (Artetxe et al., 2020) in English, German, and Turkish (App. G), conservative surface rules transfer: P1 retains 97.6–99.2% of Qwen3.5-9B's F1 in all three languages. The advantage of structured over random reduction is language-dependent: in English, S2 lies above random deletion at a similar rate for all three evaluated models, but not consistently in German and Turkish. Morphological and syntactic rules do not: M1 retains 58.5% of F1 in English but only 26.4% in German and 10.4% in Turkish, and the D-series rules remove almost the entire German context (95.8–100%), presumably because the retained dependency labels follow the English annotation scheme. Language-specific resources and validation are therefore required.

4.6 Implications

(i) Structure matters beyond length: punctuation, single-character subwords, and inflection are largely redundant, and the subject–verb–object core is the most informative aggressive reduction; rarity-based filtering is not. *(ii) Context regime determines utility:* when inputs fit into the window, conservative pruning is preferable; when truncation binds, stronger reduction can recover coverage and match or exceed the dense input, so reduction results are only interpretable together with the context budget. *(iii) Efficiency is prefill-centric* and translates into end-to-end speed-ups only when generation length is unaffected. *(iv) Transfer is structured, not universal:* broad rankings transfer, but benchmark-, language-, and parser-specific failure modes remain.

5 Conclusion

This work characterizes how LLMs respond to fixed linguistic reductions applied before tokenization. Across 15 interventions, nine models, and six benchmarks, punctuation and inflection are largely redundant, the subject–verb–object core is the most informative aggressive reduction, and frequency-based filtering falls below the reduction-adjusted trend; broad rankings transfer across architectures and tokenizers, but remain benchmark- and language-dependent.

Our central finding is the *coverage paradox* and its controlled characterization for static rules. When dense inputs exceed the context length, pruning changes not only how much information is removed but also which evidence remains accessible. On the longest LongBench-v2 documents the pruning penalty disappears and several interventions score above the unpruned baseline, whereas the same interventions clearly harm inputs that fit into the window; across models on identical items, this reversal grows with baseline truncation, and linguistic selection adds to it beyond random deletion. Input reduction should therefore be analyzed jointly with the task and the model's context budget, not through token count alone.

Limitations and future work. Our rules and thresholds are fixed by design rather than learned, so a learned compressor retains more utility at the same reduction (App. H); several rules depend on language-specific parsers, and aggressive reduction can remove negation, entities, or numbers. The random controls use a single seed, so our structural conclusions rest on the reduction-adjusted analysis, which does not use them. Our coverage evidence combines length stratification with variation in context budget across models on identical items; a controlled context-length sweep within a single model is the natural next test, and individual long-document gains rarely reach significance. All tasks focus on question answering, and the reasoning model was evaluated on subsets. The results point toward *linguistically aware tokenization* and budget-adaptive pipelines that keep the syntactic nucleus intact and apply reduction exactly where truncation would otherwise discard evidence.

Ethics statement

This work studies deterministic input reduction for publicly available LLMs on publicly available benchmarks; it involves no human subjects and no personal data. Pruning can remove qualifiers such as negations and thereby change the meaning of an input, and pruned inputs could interact differently with safety mechanisms that are tuned on fluent text. We discuss these risks, together with the potential benefits of reduced inference cost, in Appendix A.

Reproducibility statement

All models and benchmarks are publicly available; versions, licenses, sample counts, and the context budget per model are listed in Appendix B. The interventions are fully specified in Section 4.2 and Table 4, and the metrics and statistical tests in Appendix B.4. Absolute results for every model–benchmark–intervention cell are reported in Appendix C.

The supplementary material contains a self-contained reproducibility archive: the aggregate result tables for all models, benchmarks, and interventions (task and efficiency metrics, paired-test results with confidence intervals and Benjamini–Hochberg q-values, truncation summaries, cross-model rank correlations, and a completeness audit), together with a CPU-only notebook that recomputes the macro averages, the reduction-adjusted analysis, the length-stratified coverage results, the transfer statistics, and the significance counts, and checks them against the values reported in this paper. The completeness audit reports two failed generations out of about 3.97M model calls (one each for the two Llama models); all other cells are complete. The archive also records the random seed, the software environment (Python, PyTorch, CUDA, `transformers`, `spaCy` and its parser models, Stanza, Morfessor, `wordfreq`), and the LLMLingua-2 checkpoint. Full end-to-end reproduction of inference additionally requires the model weights, the benchmark datasets, and the per-sample prediction ledger, which is too large for the supplementary file; code for all interventions and evaluations will be released upon publication.

AI Assistance Statement

LLMs were used solely as writing aids to assist with grammar polishing, improving prose readability, and shortening text to fit page limits. All core conceptual ideas, technical formulations, and empirical analyses were conceived, verified, and finalized by the human authors, who take full responsibility for the contents of this manuscript.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4623–4637, 2020. doi: 10.18653/v1/2020.acl-main.421.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

Yonatan Belinkov and Yonatan Bisk. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations (ICLR)*, 2018.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster, 2023. URL `https://arxiv.org/abs/2210.09461`.

Kaj Bostrom and Greg Durrett. Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4617–4624, 2020. doi: 10.18653/v1/2020.findings-emnlp.414.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 4599–4610, 2021. doi: 10.18653/v1/2021.naacl-main.365.

Geoffrey B Duggan and Stephen J Payne. Text skimming: The process and effectiveness of foraging through text under time pressure. *Journal of experimental psychology: Applied*, 15(3):228, 2009.

Geoffrey B. Duggan and Stephen J. Payne. Skim reading by satisficing: evidence from eye tracking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pp. 1141–1150, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450302289. doi: 10.1145/1978942.1979114.

Mohsen Fayyaz, Soroush Abbasi Koohpayegani, Farnoush Rezaei Jafari, Sunando Sengupta, Hamid Reza Vaezi Joze, Eric Sommerlade, Hamed Pirsiavash, and Jürgen Gall. Adaptive token sampling for efficient vision transformers. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 396–414, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-20083-0.

Gemma Fitzsimmons, Lewis T. Jayes, Mark J. Weal, and Denis Drieghe. The impact of skim reading and navigation when reading hyperlinks on the web. *PLOS ONE*, 15(9):1–23, 09 2020. doi: 10.1371/journal.pone.0239134.

Tao Ge, Jing Hu, Lei Wang, Xun Wang, Si-Qing Chen, and Furu Wei. In-context autoencoder for context compression in a large language model. In *International Conference on Learning Representations (ICLR)*, 2024.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*, 2019.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength natural language processing in Python, 2020.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LLMLingua: Compressing prompts for accelerated inference of large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13358–13376, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.825. URL https://aclanthology.org/2023.emnlp-main.825/.

Huiqiang Jiang, Yucheng Li, Chengruidong Zhang, Qianhui Wu, Xufang Luo, Surin Ahn, Zhenhua Han, Amir H. Abdi, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. MInference 1.0: Accelerating pre-filling for long-context LLMs via dynamic sparse attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024a.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1658–1677, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.91. URL https://aclanthology.org/2024.acl-long.91/.

Marcel A Just and Patricia A Carpenter. A capacity theory of comprehension: individual differences in working memory. *Psychological review*, 99(1):122, 1992.

Zhenglun Kong, Peiyan Dong, Xiaolong Ma, Xin Meng, Wei Niu, Mengshu Sun, Xuan Shen, Geng Yuan, Bin Ren, Hao Tang, Minghai Qin, and Yanzhi Wang. Spvit: Enabling faster vision transformers via latency-aware soft token pruning. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 620–640, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-20083-0.

Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 66–71, 2018. doi: 10.18653/v1/D18-2012.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. LooGLE: Can long-context language models understand long contexts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024a.

Yucheng Li, Bo Dong, Frank Guerin, and Chenghua Lin. Compressing context to enhance inference efficiency of large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6342–6353, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.391.

Zongqian Li, Yixuan Su, and Nigel Collier. 500xcompressor: Generalized prompt compression for large language models, 2024b. URL https://arxiv.org/abs/2408.03094.

Jianfei Ma, Zhaoxin Feng, Emmanuele Chersoni, and Si Chen. Every time i hire a linguist, inference costs go down: On linguistic rules as effective prompt compressors, 2026. URL https://arxiv.org/abs/2607.25335.

Junzhu Mao, Yang Shen, Jinyang Guo, Yazhou Yao, Xiansheng Hua, and Hengtao Shen. Prune and merge: Efficient token compression for vision transformer with spatial information preserved. *IEEE Transactions on Multimedia*, 27:4670–4683, 2025. doi: 10.1109/TMM.2025.3535405.

Michael E Masson. Cognitive processes in skimming stories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8(5):400, 1982.

Meta. Llama-3.1-8b-instruct. https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct, 2024a.

Meta. Llama-3.2-3b-instruct. https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct, 2024b.

Microsoft Research. Phi-4. https://huggingface.co/microsoft/phi-4, 2024.

Mistral AI. Mistral-7b-instruct-v0.3. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3, 2024.

Fabio Montello, Ronja Güldenring, and Lazaros Nalpantidis. Clustvit: Clustering-based token merging for semantic segmentation, 2025. URL https://arxiv.org/abs/2510.01948.

Jesse Mu, Xiang Li, and Noah Goodman. Learning to compress prompts with gist tokens. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 19327–19352. Curran Associates, Inc., 2023.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Dongmei Zhang. LLMLingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 963–981, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.57.

Peter Pirolli and Stuart Card. Information foraging. *Psychological review*, 106(4):643, 1999.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 101–108, 2020. doi: 10.18653/v1/2020.acl-demos.14.

Qwen Team. Qwen2.5-14b-instruct. https://huggingface.co/Qwen/Qwen2.5-14B-Instruct, 2024a.

Qwen Team. Qwen2.5-7b-instruct. https://huggingface.co/Qwen/Qwen2.5-7B-Instruct, 2024b.

Qwen Team. Qwen3-4b-thinking-2507. https://huggingface.co/Qwen/Qwen3-4B-Thinking-2507, 2025.

Qwen Team. Qwen3.5-4b. https://huggingface.co/Qwen/Qwen3.5-4B, 2026a.

Qwen Team. Qwen3.5-9b. https://huggingface.co/Qwen/Qwen3.5-9B, 2026b.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 784–789, 2018. doi: 10.18653/v1/P18-2124.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 13937–13949. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/747d3443e319a22747fbb873e8b2f9f2-Paper.pdf.

Keith Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372, 1998.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020. doi: 10.1162/tacl_a_00349.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1715–1725, 2016. doi: 10.18653/v1/P16-1162.

Robyn Speer. rspeer/wordfreq: v3.0, 9 2022. URL https://doi.org/10.5281/zenodo.7199437.

Yutao Sun, Li Dong, Yi Zhu, Shaohan Huang, Wenhui Wang, Shuming Ma, Quanlu Zhang, Jianyong Wang, and Furu Wei. You only cache once: Decoder-decoder architectures for language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 7339–7361. Curran Associates, Inc., 2024. doi: 10.52202/079017-0235.

John Sweller, Jeroen J. G. van Merriënboer, and Fred Paas. Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, 31(2):261–292, June 2019. ISSN 1573-336X. doi: 10.1007/s10648-019-09465-5.

Xudong Tan, Peng Ye, Chongjun Tu, Jianjian Cao, Yaoxin Yang, Lin Zhang, Dongzhan Zhou, and Tao Chen. Tokencarve: Information-preserving visual token compression in multimodal large language models, 2025. URL https://arxiv.org/abs/2503.10501.

Hoai-Chau Tran, Duy M. H. Nguyen, Duy M. Nguyen, TrungTin Nguyen, Ngan Le, Pengtao Xie, Daniel Sonntag, James Zou, Binh T. Nguyen, and Mathias Niepert. Accelerating transformers with spectrum-preserving token merging. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 30772–30810. Curran Associates, Inc., 2024. doi: 10.52202/079017-0969.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 5998–6008, 2017.

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, and Mikko Kurimo. Morfessor 2.0: Python implementation and extensions for Morfessor Baseline. Aalto University publication series SCIENCE + TECHNOLOGY 25/2013, Aalto University, 2013.

Elena Voita, Rico Sennrich, and Ivan Titov. Analyzing the source and target contributions to predictions in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pp. 1126–1140, 2021. doi: 10.18653/v1/2021.acl-long.91.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 37, 2024.

Weihao Ye, Qiong Wu, Wenhao Lin, and Yiyi Zhou. Fit and prune: Fast and training-free visual token pruning for multi-modal large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(21):22128–22136, Apr. 2025. doi: 10.1609/aaai.v39i21.34366. URL https://ojs.aaai.org/index.php/AAAI/article/view/34366.

Xin Yuan, Hongliang Fei, and Jinoo Baek. Efficient transformer adaptation with soft token merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 3658–3668, June 2024.

Qizhe Zhang, Aosong Cheng, Ming Lu, Renrui Zhang, Zhiyong Zhuo, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. Beyond text-visual attention: Exploiting visual cues for effective token pruning in vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20857–20867, October 2025.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang "Atlas" Wang, and Beidi Chen. H2o: Heavy-hitter oracle for efficient generative inference of large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34661–34710. Curran Associates, Inc., 2023.

APPENDIX

This appendix provides the experimental details, complete results, and additional analyses underlying the main text.

- **Appendix A** discusses broader impact and potential risks.
- **Appendix B** specifies models, benchmarks, pruning patterns, metrics, statistical tests, compute resources, and the computational overhead of the interventions.
- **Appendix C** reports aggregate and per-model results, the sensitivity of the composite score to accuracy weighting, cross-model ranking stability, significance tests, the reasoning model, and absolute results for every model–benchmark–intervention cell.
- **Appendix D** provides the full long-context analysis underlying the coverage paradox (truncation incidence and LongBench-v2 length categories for all models).
- **Appendix E** analyzes peak GPU memory, and **Appendix F** summarizes all relative quality and efficiency metrics per model.
- **Appendix G** extends the characterization to German and Turkish.
- **Appendix H** reports LLMLingua-2 as an out-of-scope reference for a *learned* compressor.
- **Appendix I** compares the subword segmentation of the evaluated tokenizers.

A BROADER IMPACT

This paper presents an empirical characterization of word- and subword-level pruning of LLM inputs using deterministic rules that require no LLM-based scoring and no task-specific training. Its primary goal is to understand which linguistic structures LLMs rely on and how input reduction interacts with finite context capacity.

Potential benefits. Reducing input length before tokenization lowers inference cost, latency, and memory, which may support more sustainable deployment and improve accessibility in resource-constrained settings. Because the interventions require no training, auxiliary models, or learned components, they introduce no additional data dependencies. The coverage paradox further indicates that, for documents exceeding the context length, reduction can make more of the source document accessible to the model instead of silently discarding its tail, which is relevant for document-heavy domains such as law, medicine, and science.

Potential risks. Although the evaluated models are fairly robust to the loss of function words, removing modifiers such as “not”, “rarely”, or “unless” can invert the meaning of a statement, and aggressive rules can remove entities or numbers. Pruned inputs should therefore not be used in high-stakes settings without task-specific validation. Moreover, safety classifiers that are tuned on fluent text may behave differently on pruned inputs, which could in principle be exploited to evade input filters. Finally, efficiency improvements can indirectly increase the scale of LLM deployment and thereby amplify existing risks such as misuse, over-reliance, or biased outputs; these risks are not specific to this work.

B EXPERIMENTAL DETAILS

B.1 MODELS

Table 2 lists all evaluated models. All models were run with the Hugging Face `transformers` library using the default configuration and generation settings provided by the original authors; no hyper-parameters or weights were modified. Each model is evaluated with the native context length specified in its configuration, except for Qwen3.5-4B and Qwen3.5-9B, which we cap at 32,768 tokens. The prompt budget equals this limit minus a benchmark-specific answer budget (8 tokens for LongBench-v2, MMLU-Pro, and ARC-Challenge; 32 for SQuAD v2 and XQuAD; 64 for QASPER; 96 for LooGLE), minus an additional reserve of 8,192 tokens for the reasoning trace of Qwen3-4B-Thinking-2507. If the context does not fit into the remaining budget after accounting for instructions, question, options, and a 32-token safety margin, it is truncated at the token level with the target model's tokenizer by keeping approximately equal parts from the beginning and the end of the context, separated by the marker `...[context truncated]...`; instructions, question, and answer options are never truncated. The reasoning model is evaluated on a deterministic, stratified subset of $\min(100, \lfloor 0.1 \cdot N \rfloor)$ items per benchmark (50 on LongBench-v2), fixed across all interventions. FlashAttention-2 was enabled for all compatible models.

Table 2: **Evaluated pre-trained models and licenses.** †Evaluated on fixed subsets (100 samples per benchmark, 50 on LongBench-v2).

Model	Type	License	Source (Hugging Face)
Llama-3.2-3B-Instruct (Meta, 2024b)	Instruct	Llama 3.2 Community License	`https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct`
Llama-3.1-8B-Instruct (Meta, 2024a)	Instruct	Llama 3.1 Community License	`https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct`
Mistral-7B-Instruct-v0.3 (Mistral AI, 2024)	Instruct	Apache-2.0	`https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3`
Qwen2.5-7B-Instruct (Qwen Team, 2024b)	Instruct	Apache-2.0	`https://huggingface.co/Qwen/Qwen2.5-7B-Instruct`
Qwen2.5-14B-Instruct (Qwen Team, 2024a)	Instruct	Apache-2.0	`https://huggingface.co/Qwen/Qwen2.5-14B-Instruct`
Qwen3.5-4B (Qwen Team, 2026a)	Instruct	Apache-2.0	`https://huggingface.co/Qwen/Qwen3.5-4B`
Qwen3.5-9B (Qwen Team, 2026b)	Instruct	Apache-2.0	`https://huggingface.co/Qwen/Qwen3.5-9B`
Phi-4 (Microsoft Research, 2024)	Instruct	MIT	`https://huggingface.co/microsoft/phi-4`
Qwen3-4B-Thinking-2507† (Qwen Team, 2025)	Reasoning	Apache-2.0	`https://huggingface.co/Qwen/Qwen3-4B-Thinking-2507`

B.2 BENCHMARKS

Table 3 summarizes the six main benchmarks. QASPER is evaluated on 1,451 question–answer pairs over its test papers. LongBench-v2 additionally provides a length category for each item (short, medium, long), which we use to stratify the long-context analysis (App. D). The multilingual study uses XQuAD (Artetxe et al., 2020) (App. G).

B.3 PRUNING PATTERNS

Table 4 defines all interventions, and Table 5 applies each of them to the same passage. The dependency- and POS-based rules use the English `spaCy` pipeline (Honnibal et al., 2020); for the multilingual study we use language-specific `spaCy` pipelines and, for Turkish dependency parsing, Stanza (Qi et al., 2020). Morphological roots are estimated with Morfessor (Virpioja et al., 2013). For the multilingual study, an additional random reference R15 matches the lower reduction rates of conservative rules in German and Turkish.

Table 3: **Benchmark specifications.** Samples refer to the full evaluation sets used for the non-reasoning models.

Benchmark	Split	Metric	License	Samples	URL
SQuAD v2 (Rajpurkar et al., 2018)	Validation	F1, EM	CC BY-SA 4.0	11,873	https://hf.co/datasets/rajpurkar/squad_v2
LongBench-v2 (Bai et al., 2025)	Train	Accuracy	Apache-2.0	503	https://hf.co/datasets/zai-org/LongBench-v2
LooGLE LongDepQA (Li et al., 2024a)	Test	ROUGE-L	CC BY-SA 4.0	1,101	https://hf.co/datasets/bigai-nlco/LooGLE
QASPER (Dasigi et al., 2021)	Test	F1, EM	CC BY 4.0	1,451	https://hf.co/datasets/allenai/qasper
MMLU-Pro (Wang et al., 2024)	Test	Accuracy	MIT	12,032	https://hf.co/datasets/TIGER-Lab/MMLU-Pro
ARC-Challenge (Clark et al., 2018)	Test	Accuracy	CC BY-SA 4.0	1,172	https://hf.co/datasets/allenai/ai2_arc
XQuAD (en, de, tr) (Artetxe et al., 2020)	Test	F1	CC BY-SA 4.0	1,190 per lang.	https://hf.co/datasets/google/xquad

Table 4: **Pruning patterns used in this work.** Each rule is static, deterministic, and applied to the raw string prior to tokenization.

Pattern	Description
S1	For each whitespace-delimited word, retain only the subword with the highest count of alphabetic characters, approximating the semantic core of the word (e.g., `un-believ-able` → `believ`).
S2	Remove subwords whose alphanumeric core has length one, eliminating isolated characters and punctuation fragments.
P1	Replace all non-alphanumeric characters by whitespace (regex `[^\w\s]`) and normalize whitespace; preserves only words and numbers.
L1	Using `spaCy`, remove stopwords, punctuation, and spaces; retains content words.
L2	Using `spaCy` part-of-speech (POS) tags, remove function words (determiners, adpositions, conjunctions, particles), punctuation, and spaces.
F1	Remove each word whose alphanumeric core has an English Zipf frequency above 4 (`wordfreq`), i.e., highly frequent, low-information items.
F2	As F1, with Zipf threshold 3 (retains only words that occur fewer than about once per million words).
D1	Using dependency parses, retain only words that are syntactic arguments (subjects and objects).
D2	Retain only subjects, the main predicate (`ROOT`), and objects, i.e., a minimal subject–verb–object skeleton.
M1	Estimate a morphological root with Morfessor and retain the subword most similar to it.
M2	Replace each word by its estimated stem; tests robustness to inflectional and derivational variation.
R20–R80	Delete 20, 40, 60, or 80% of the subword tokens of the input uniformly at random with a fixed seed (42); controls for sequence-length reduction without linguistic bias. Because deletion operates on tokens, retained words can be fragmentary (e.g., `2025` → `025`). R15 is used additionally in the multilingual study.

B.4 Metrics and statistical testing

For every model, benchmark, and intervention, we report the benchmark's primary metric (accuracy for LongBench-v2, MMLU-Pro, and ARC-Challenge; token-level F1 for SQuAD v2, QASPER, and XQuAD; ROUGE-L for LooGLE) and the relative quality Q_{rel}, i.e., the ratio of the pruned to the unpruned score of the same model on the same items. Model-level summaries are macro means over the six benchmarks, weighting each benchmark equally. The pruning ratio ρ is computed on prompt tokens with the tokenizer of the evaluated model. The end-to-end speed-up is the ratio of baseline to pruned task completion time t_{TC} measured with a warm cache, including the time spent on pruning; the t_{TTFT} speed-up is defined analogously. Memory reduction is $1 - M_{GPU}^{\text{pruned}} / M_{GPU}^{\text{base}}$, where M_{GPU}

Table 5: **Qualitative examples of one passage after applying each pruning pattern.** The last column reports the character-level reduction for this passage. Random references delete subword tokens, so fragments of words and numbers (e.g., `dney`, `025`) are expected; subword rules (S1, M1) can produce fragments for the same reason.

Pattern	Pruned Version	Reduction
—	Sydney is the capital city of the state of New South Wales, and the most populous city in Australia. Greater Sydney consists of 658 suburbs, spread across 33 local government areas. Residents of the city are colloquially known as "Sydneysiders". The estimated population in June 2025 is about 66% of Australia's population. The city's nicknames include the Emerald City and the Harbour City.	—
S1	dney is the capital city of the state of New South Wales and the most populous city in Australia Greater Sydney consists of suburbs spread across local government areas Residents of the city are collo known as iders The estimated population in June is about of Australia population The city nick include the Emerald City and the Harbour City	-34.78%
S2	Sydney is the capital city of the state of New South Wales and the most populous city in Australia Greater Sydney consists of suburbs spread across local government areas Residents of the city are colloquially known asSyneysiders The estimated population in June is about of Australia population The city nicknames include the Emerald City and the Harbour City	-29.34%
P1	Sydney is the capital city of the state of New South Wales and the most populous city in Australia Greater Sydney consists of 658 suburbs spread across 33 local government areas Residents of the city are colloquially known as Sydneysiders The estimated population in June 2025 is about 66 of Australia s population The city s nicknames include the Emerald City and the Harbour City	-9.78%
L1	Sydney capital city state New South Wales populous city Australia Greater Sydney consists 658 suburbs spread 33 local government areas Residents city colloquially known Sydneysiders estimated population June 2025 66 Australia population city nicknames include Emerald City Harbour City	-38.04%
L2	Sydney is capital city state New South Wales most populous city Australia Greater Sydney consists 658 suburbs spread 33 local government areas Residents city are colloquially known Sydneysiders estimated population June 2025 is about 66 Australia population city nicknames include Emerald City Harbour City	-32.60%
F1	populous 658 suburbs, 33 colloquially "Sydneysiders". 2025 66% Australia's nicknames Emerald Harbour	-61.95%
F2	658 33 colloquially "Sydneysiders". 2025 66%	-73.91%
D1	Sydney state Wales Australia Sydney suburbs areas Residents city Sydneysiders population June population nicknames City	-78.26%
D2	Sydney is Sydney consists Residents known population is nicknames include City	-85.86%
M1	dney is the capital city of the state of New South Wales and the most populous city in Australia Greater Sydney consists of suburbs spread across local government areas Residents of the city are collo known as neys The estimated population in June is about of Australia population The city nick include the Emerald City and the Harbour City	-34.78%
M2	sydney is the capit citi of the state of new south wale and the most popul citi in australia greater sydney consist of suburb spread across local govern area resid of the citi are colloqui known as sydneysid the estim popul in june is about of australia popul the citi nicknam includ the emerald citi and the harbour citi	-18.47%
R20	Sydney city of the state of New Wales, and the most city in Australia Greater Sydney consists of 58 suburbs, across33 local government areas. Residents of city colloquially known as "Sydneysiders". The estimated population in June 025 is about 66% of's The citynames include the City and the Harbour City.	-19.56%
R40	Sydney is the capital of the state of New South Wales, and the most populous Australia. Greater Sydney65 suburbs, across 3 local government areas. Residents citydiders The in June02 about 66% Australia The's nick include and the Harbour City.	-39.13%
R60	dney the state New and the most city in Greater Sydney consists of across areas. Residents city are known asdiders in June 5 about6 of Australia's. The include	-61.95%

denotes the peak GPU allocation above the static footprint of the loaded model. Truncation incidence is the fraction of inputs whose tokenized prompt exceeds the configured context length.

To test differences to the baseline, we compare per-sample scores of each intervention with those of the unpruned baseline on the same items, using a two-sided exact McNemar test for binary accuracy (LongBench-v2, MMLU-Pro, ARC-Challenge) and a two-sided paired sign-flip permutation test with 10,000 resamples for continuous scores (F1 for SQuAD v2, QASPER, and XQuAD; ROUGE-L for LooGLE). Effect uncertainty is estimated with paired bootstrap confidence intervals (10,000 resamples), and the false discovery rate is controlled with the Benjamini–Hochberg procedure at $q \leq 0.05$ within each model, benchmark, and analysis family.

B.5 Compute resources and pruning overhead

All experiments were conducted on an Azure `Standard_NC80adis_H100_v5` instance managed via Databricks (Databricks Runtime 17.3 LTS), equipped with two NVIDIA H100 NVL (94 GB VRAM each), 40 physical CPU cores (80 vCPUs), and 640 GB of system RAM (Python 3.12, PyTorch 2.7.0, CUDA 12.6, `transformers` 4.51.3, FlashAttention 2.7.4). The software environment is documented in full in the supplementary archive.

Table 6 reports the average time required to apply each pattern to an input of 41,959 characters (the mean context length across benchmarks). The overhead is at most about 0.5 s and is included in all reported end-to-end speed-ups. The higher cost of L2, D1, and D2 stems from full `spaCy` POS tagging and dependency parsing in our research pipeline; production implementations could replace parsing by precomputed lookup tables for the lexical rules, although the syntactic rules intrinsically require a parser.

Table 6: **Average computational overhead of the pruning patterns** (milliseconds per input of 41,959 characters; the patterns do not depend on the evaluated LLM).

Pattern	ms	Pattern	ms	Pattern	ms
S1	11.7	L2	244.5	D2	498.6
S2	19.2	F1	5.1	M1	19.2
P1	1.1	F2	4.5	M2	2.3
L1	1.4	D1	498.8	R20–R80	14.2–14.5

B.6 Reproducibility details

Pruning scope. Table 7 lists the prompt fields to which the interventions are applied. The actual fields affected for every sample are logged with each result record.

Table 7: **Prompt fields affected by pruning.**

Benchmark	Context	Question	Answer options
SQuAD v2, QASPER, LooGLE, XQuAD	✓	✓	–
LongBench-v2	✓	✓	✓
MMLU-Pro, ARC-Challenge	(none)	✓	✗
LLMLingua-2 (Appendix H)	✓	✗	✗

Random references. Random deletion operates on subword tokens with a deterministic, item-specific seed derived from the global seed (42), so deletion masks differ across items and item-level tests average over thousands of distinct masks; however, each item is evaluated under a single mask, so mask-level variance per item is not quantified. For this reason, comparisons with random deletion are reported as single-seed references, and our structural conclusions rely primarily on the reduction-adjusted analysis (Tab. 13), which is fitted to the linguistic interventions only and does not use the random references.

Implementation details. Exact model revisions, decoding settings, prompt templates, answer-extraction rules, and the segmentation of long documents for parsing are part of the released code upon publication. Reported completion times are warm-cache times that include the pruning overhead. The subsets for Qwen3-4B-Thinking-2507 are fixed across interventions, and their item indices are released with the code.

Software versions. spaCy 3.8.14 (`en_core_web_sm` 3.8.0, `de_core_news_sm` 3.8.0), Stanza 1.14.0 (Turkish `tr/default`), Morfessor 2.0.6, `wordfreq` 3.1.1, global seed 42.

C Additional Empirical Results

C.1 Baselines and aggregate results

Table 8 reports the unpruned baseline of every model–benchmark cell together with its truncation incidence on the long-context benchmarks. Table 9 aggregates all interventions over the 54 model–benchmark cells. Averaged over all cells, P1 is the only intervention without a loss (Q_{rel}=1.003), and D2 attains the highest composite score (S_c=0.432) at the highest reduction (ρ=0.683), the largest speed-up (1.96× among linguistic rules), and the lowest truncation incidence. R80 attains a slightly larger mean speed-up (2.05×) but lower quality at a lower reduction. The frequency rules F1 and F2 lie below D1, D2, and R80 in S_c.

Table 8: **Unpruned baseline scores** (primary metric ×100: F1 for SQuAD v2 and QASPER, ROUGE-L for LooGLE, accuracy otherwise) and fraction of baseline inputs that exceed the model's configured context length and are therefore truncated (in parentheses, long-context benchmarks only). †Evaluated on subsets (100 samples per benchmark; 50 on LongBench-v2).

Model	SQuAD v2	LongBench-v2	LooGLE	QASPER	MMLU-Pro	ARC-C
Llama-3.2-3B	41.16	22.27 (36.2%)	11.24 (0.0%)	33.71	28.19	69.11
Llama-3.1-8B	62.13	28.23 (36.2%)	14.65 (0.0%)	38.30	36.18	78.24
Mistral-7B	41.82	25.05 (80.3%)	14.83 (58.1%)	31.87	29.75	73.29
Qwen2.5-7B	72.21	30.22 (76.9%)	10.81 (50.9%)	43.21	41.79	89.42
Qwen2.5-14B	74.39	34.79 (76.9%)	9.91 (50.9%)	35.19	50.71	93.34
Qwen3.5-4B	62.94	35.39 (77.9%)	11.00 (53.0%)	44.79	46.73	92.24
Qwen3.5-9B	70.04	35.79 (77.9%)	12.09 (53.0%)	41.39	51.97	94.54
Phi-4	54.40	22.27 (94.4%)	7.32 (83.5%)	29.73	53.07	94.03
Qwen3-4B-Think.†	54.29	14.00 (22.0%)	22.14 (0.0%)	47.82	19.00	36.00

Table 9: **Aggregate results over all 54 model–benchmark cells** (macro means; ± denotes the descriptive 95% half-width over cells). Truncation denotes the fraction of inputs exceeding the configured context length.

Pattern	Q_{rel}	ρ	Speed-up	Mem. red.	Truncation	S_c
Baseline	1.000 ± 0.000	0.000	1.00×	0.000	0.172	0.000
P1	1.003 ± 0.018	0.121	1.06×	0.062	0.133	0.121
S1	0.862 ± 0.031	0.248	1.07×	0.112	0.099	0.214
S2	0.930 ± 0.025	0.196	1.07×	0.094	0.109	0.182
L1	0.761 ± 0.042	0.362	1.16×	0.140	0.098	0.275
L2	0.880 ± 0.030	0.277	1.11×	0.143	0.111	0.243
F1	0.580 ± 0.053	0.546	1.46×	0.291	0.074	0.317
F2	0.563 ± 0.055	0.639	1.72×	0.343	0.060	0.360
D1	0.612 ± 0.059	0.638	1.71×	0.389	0.033	0.390
D2	0.632 ± 0.053	0.683	1.96×	0.392	0.026	0.432
M1	0.862 ± 0.027	0.247	1.01×	0.117	0.099	0.213
M2	0.925 ± 0.029	0.172	1.05×	0.096	0.116	0.159
R20	0.883 ± 0.025	0.176	1.08×	0.077	0.125	0.156
R40	0.753 ± 0.030	0.336	1.23×	0.166	0.100	0.253
R60	0.649 ± 0.036	0.495	1.51×	0.268	0.074	0.321
R80	0.580 ± 0.050	0.654	2.05×	0.374	0.042	0.379

C.2 Per-model results

Tables 10–12 report the macro-averaged relative quality, pruning ratio, and end-to-end speed-up of every intervention for every model, and Figures 5–7 visualize the corresponding performance–efficiency landscapes. Absolute per-benchmark results for each model are listed in Tables 22–30 at the end of this section.

Consistent hierarchy. Across models, relative quality decreases with the pruning ratio, but the rate of decrease depends on what is removed. The conservative patterns P1, S2, and M2 retain 85–104% of baseline quality at 12–25% reduction. Among aggressive patterns, D2 attains the highest composite score on every model and, in our single-seed random reference, exceeds R80 in both relative quality and reduction on every model. F2 retains less quality than R80 on six of nine models despite removing fewer tokens, and L1 and R40 lie close to each other, indicating that stopword

removal and random deletion at a similar rate are similarly harmful when averaged over tasks. Phi-4 is the most robust model overall (D2: Q_{rel}=0.745), in part because of its large gains on truncated LongBench-v2 inputs (App. D).

Efficiency. Speed-ups scale with the pruning ratio and are largest for long inputs, where prefill dominates. Because the TTFT gains translate into end-to-end gains only if generation length is unaffected, end-to-end speed-ups vary across models: they are largest for the Llama models (D2: 2.75× and 2.92×) and smallest among the non-reasoning models for Mistral-7B (0.96×) and Phi-4 (1.31×), whose completion time on LongBench-v2 increases under several interventions (Tabs. 24 and 29). On MMLU-Pro and ARC-Challenge, whose inputs are short, speed-ups remain close to one for all models.

Table 10: **Relative task quality (Q_{rel}) per model and intervention**, macro-averaged over the six benchmarks. Values above 1 indicate that the pruned input scores higher than the unpruned baseline. †Evaluated on subsets (100 samples per benchmark; 50 on LongBench-v2).

Model	P1	S1	S2	L1	L2	F1	F2	D1	D2	M1	M2	R20	R40	R60	R80
Llama-3.2-3B	1.009	0.880	0.959	0.714	0.821	0.602	0.600	0.570	0.601	0.880	0.955	0.871	0.769	0.642	0.592
Llama-3.1-8B	0.978	0.864	0.926	0.729	0.849	0.589	0.536	0.568	0.614	0.864	0.907	0.885	0.772	0.658	0.582
Mistral-7B	1.021	0.838	0.951	0.789	0.892	0.594	0.557	0.620	0.660	0.840	0.961	0.907	0.789	0.665	0.594
Qwen2.5-7B	1.009	0.856	0.933	0.773	0.888	0.629	0.597	0.598	0.618	0.852	0.914	0.892	0.742	0.640	0.583
Qwen2.5-14B	1.016	0.923	0.970	0.794	0.909	0.628	0.593	0.608	0.633	0.907	0.945	0.899	0.781	0.658	0.583
Qwen3.5-4B	0.966	0.830	0.899	0.719	0.842	0.530	0.523	0.547	0.561	0.825	0.892	0.876	0.738	0.618	0.556
Qwen3.5-9B	1.000	0.837	0.921	0.759	0.862	0.563	0.512	0.571	0.588	0.846	0.917	0.884	0.743	0.612	0.553
Phi-4	1.040	0.921	0.933	0.831	0.930	0.671	0.662	0.728	0.745	0.908	0.980	0.875	0.808	0.713	0.679
Qwen3-4B-Think.†	0.992	0.811	0.875	0.744	0.925	0.415	0.493	0.693	0.664	0.832	0.851	0.859	0.631	0.635	0.498
Mean	1.003	0.862	0.930	0.761	0.880	0.580	0.563	0.612	0.632	0.862	0.925	0.883	0.753	0.649	0.580

Table 11: **Token-level pruning ratio (ρ) per model and intervention**, macro-averaged over the six benchmarks and measured with each model's own tokenizer. †Evaluated on subsets (100 samples per benchmark; 50 on LongBench-v2).

Model	P1	S1	S2	L1	L2	F1	F2	D1	D2	M1	M2	R20	R40	R60	R80
Llama-3.2-3B	0.120	0.234	0.167	0.369	0.282	0.559	0.651	0.641	0.684	0.232	0.171	0.174	0.335	0.495	0.656
Llama-3.1-8B	0.120	0.234	0.167	0.369	0.282	0.559	0.651	0.641	0.684	0.232	0.171	0.174	0.335	0.495	0.656
Mistral-7B	0.131	0.285	0.248	0.349	0.271	0.523	0.635	0.634	0.681	0.286	0.184	0.184	0.341	0.498	0.655
Qwen2.5-7B	0.115	0.246	0.198	0.358	0.273	0.543	0.634	0.638	0.682	0.245	0.166	0.173	0.333	0.493	0.653
Qwen2.5-14B	0.115	0.246	0.198	0.358	0.273	0.543	0.634	0.638	0.682	0.245	0.166	0.173	0.333	0.493	0.653
Qwen3.5-4B	0.124	0.250	0.208	0.366	0.281	0.550	0.640	0.640	0.684	0.249	0.172	0.180	0.338	0.496	0.655
Qwen3.5-9B	0.124	0.250	0.208	0.366	0.281	0.550	0.640	0.640	0.684	0.249	0.172	0.180	0.338	0.496	0.655
Phi-4	0.120	0.234	0.168	0.369	0.281	0.558	0.651	0.641	0.684	0.232	0.171	0.174	0.335	0.495	0.656
Qwen3-4B-Think.†	0.118	0.256	0.206	0.349	0.267	0.526	0.614	0.635	0.679	0.254	0.178	0.173	0.332	0.489	0.648
Mean	0.121	0.248	0.196	0.362	0.277	0.546	0.639	0.638	0.683	0.247	0.172	0.176	0.336	0.495	0.654

Table 12: **End-to-end speed-up (baseline t_{TC} / pruned t_{TC}) per model and intervention**, macro-averaged over the six benchmarks (warm cache, pruning overhead included). Values below 1 indicate slower completion, typically caused by longer generated outputs. †Evaluated on subsets (100 samples per benchmark; 50 on LongBench-v2).

Model	P1	S1	S2	L1	L2	F1	F2	D1	D2	M1	M2	R20	R40	R60	R80
Llama-3.2-3B	1.10	1.14	1.12	1.56	1.31	2.24	2.71	2.72	2.92	1.04	1.13	1.15	1.44	1.97	2.91
Llama-3.1-8B	1.10	1.24	1.15	1.37	1.30	1.77	2.02	2.35	2.75	1.23	1.13	1.17	1.42	1.88	2.71
Mistral-7B	1.06	1.00	1.06	0.84	0.84	1.16	1.20	0.87	0.96	0.89	0.92	1.07	1.18	1.32	1.51
Qwen2.5-7B	1.05	1.03	1.06	1.17	1.17	1.29	1.50	1.62	1.79	0.97	1.08	1.08	1.17	1.32	1.70
Qwen2.5-14B	1.06	1.11	1.10	1.21	1.17	1.48	1.97	2.05	2.43	1.05	1.10	1.10	1.36	1.89	2.78
Qwen3.5-4B	1.04	1.14	1.09	1.20	1.17	1.42	1.59	1.58	2.07	1.14	1.08	1.08	1.22	1.40	1.80
Qwen3.5-9B	1.07	1.08	1.09	1.29	1.18	1.61	2.01	2.07	2.28	1.00	1.12	1.14	1.34	1.67	2.22
Phi-4	1.06	0.99	1.02	0.94	0.94	1.13	1.31	1.10	1.31	0.92	0.96	1.07	1.15	1.34	1.72
Qwen3-4B-Think.†	1.01	0.88	0.95	0.86	0.91	1.02	1.19	1.04	1.13	0.85	0.93	0.86	0.82	0.83	1.12
Mean	1.06	1.07	1.07	1.16	1.11	1.46	1.72	1.71	1.96	1.01	1.05	1.08	1.23	1.51	2.05

Reduction-adjusted quality. Table 13 reports, for every model, the residual of Q_{rel} relative to the *linguistic-intervention trend*: a linear regression of Q_{rel} on ρ fitted to the 11 linguistic interventions only. The random references do not enter the fit, so the ranking of the linguistic rules does not depend on them; they are shown against the same trend for comparison. Under this specification, D2 has the largest residual on all nine models ($+0.049$ to $+0.127$) and lies above R80 on every model, P1 is among the three largest residuals on seven models, F1 lies below the trend on every model, and F2 on eight of nine. The residual is a descriptive linear adjustment, not a causal estimate. Because quality may degrade nonlinearly with reduction, we repeated the analysis with a quadratic fit: the D2 residual remains positive on all nine models ($+0.014$ to $+0.060$) and ranks first or second on eight of them, while F1 remains negative on every model. The ordering of the aggressive rules is therefore not an artifact of the linear specification, although the exact residuals are.

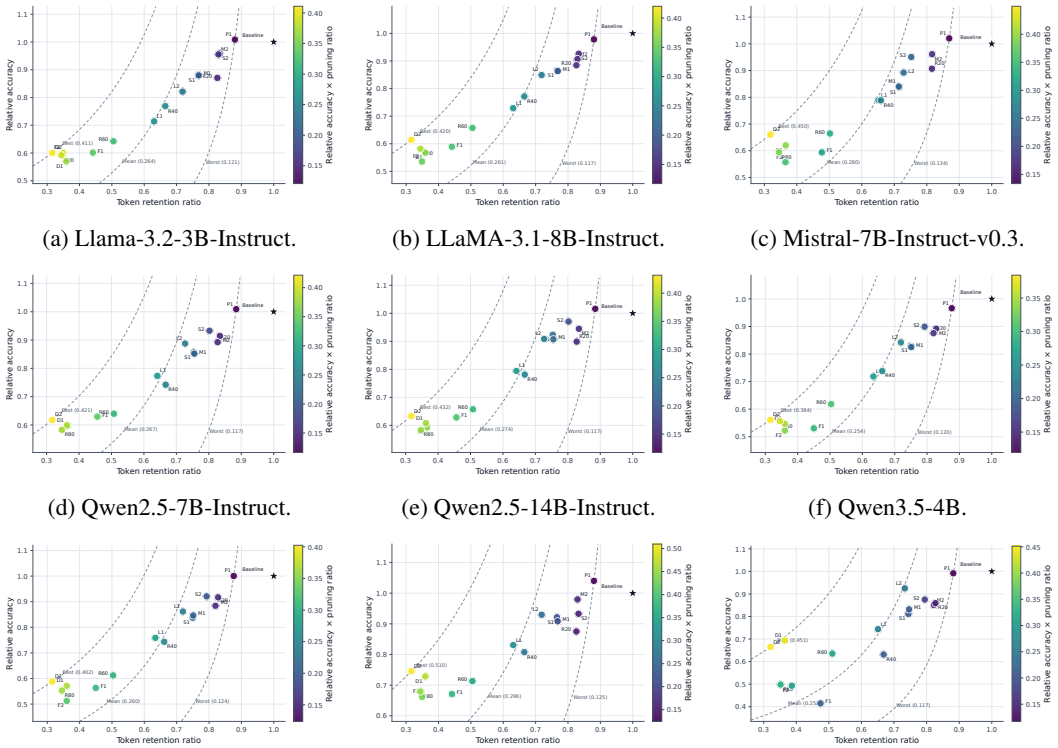
Item-weighted aggregation. Macro means weight each benchmark equally and can amplify ratios computed on low baselines. Table 14 contrasts them with item-weighted ratios. Item weighting is dominated by SQuAD v2 and MMLU-Pro, which contribute 84% of all items. With SQuAD v2 included, R80 obtains a higher item-weighted ratio than D2, because heavy pruning raises SQuAD v2 scores through additional no-answer predictions (Sec. 4.3). Without SQuAD v2, D2 again exceeds R80 (0.631 vs. 0.614). We therefore report absolute item counts wherever relative gains are large, e.g., for Phi-4 on long LongBench-v2 documents (15 $\rightarrow$ 30 of 108 items correct with D2).

Table 13: **Quality relative to the linguistic-intervention trend.** For each model, a linear regression of the macro-averaged relative quality Q_{rel} on the pruning ratio ρ is fitted to the 11 linguistic interventions only; the table reports residuals (positive: more quality retained than the linguistic trend predicts at the achieved reduction). Random references (right) are evaluated against the same trend but do not enter the fit. Bold marks the largest residual among the linguistic interventions. The last column gives the sign of the D2 residual under a quadratic fit. †Evaluated on subsets (100 samples per benchmark; 50 on LongBench-v2).

Model	P1	S1	S2	L1	L2	F1	F2	D1	D2	M1	M2	R20	R40	R60	R80	D2 (quad.)
Llama-3.2-3B	+0.037	-0.006	+0.023	-0.070	-0.029	-0.040	+0.028	-0.010	**+0.053**	-0.008	+0.022	-0.060	-0.041	-0.047	+0.023	+0.014
Llama-3.1-8B	+0.027	-0.004	+0.010	-0.040	+0.016	-0.041	-0.027	-0.003	**+0.075**	-0.006	-0.007	-0.027	-0.022	-0.019	+0.022	+0.052
Mistral-7B	+0.031	-0.032	+0.052	-0.032	+0.012	-0.091	-0.041	+0.021	**+0.098**	-0.029	+0.012	-0.042	-0.038	-0.040	+0.011	+0.058
Qwen2.5-7B	+0.039	-0.021	+0.021	-0.025	+0.029	-0.038	-0.007	-0.002	**+0.049**	-0.027	-0.020	-0.037	-0.074	-0.063	-0.007	+0.026
Qwen2.5-14B	+0.007	+0.015	+0.025	-0.028	+0.021	-0.052	-0.018	-0.000	**+0.059**	-0.003	-0.025	-0.066	-0.061	-0.062	-0.014	+0.043
Qwen3.5-4B	+0.026	-0.011	+0.025	-0.031	+0.026	-0.074	-0.011	+0.014	**+0.063**	-0.017	-0.010	-0.020	-0.033	-0.028	+0.035	+0.036
Qwen3.5-9B	+0.034	-0.027	+0.022	-0.013	+0.022	-0.060	-0.040	+0.020	**+0.072**	-0.019	-0.010	-0.037	-0.050	-0.055	+0.013	+0.046
Phi-4	+0.047	-0.007	-0.033	-0.022	+0.028	-0.075	-0.032	+0.029	**+0.070**	-0.021	+0.015	-0.087	-0.064	-0.068	-0.012	+0.045
Qwen3-4B-Think.†	+0.060	-0.024	+0.005	-0.025	+0.098	-0.231	-0.090	+0.125	**+0.127**	-0.005	-0.039	-0.034	-0.150	-0.035	-0.061	+0.060

Table 14: **Macro- versus item-weighted (micro) aggregation** over all model–benchmark cells. The micro ratio is the pooled pruned score divided by the pooled baseline score, weighting every evaluated item equally; it is dominated by SQuAD v2 and MMLU-Pro, which contribute 84% of all items.

	P1	F2	D1	D2	R60	R80
Macro mean Q_{rel}	1.003	0.563	0.612	0.632	0.649	0.580
Micro ratio, all benchmarks	0.983	0.586	0.600	0.618	0.670	0.655
Micro ratio, without SQuAD v2	0.983	0.613	0.648	0.631	0.694	0.614

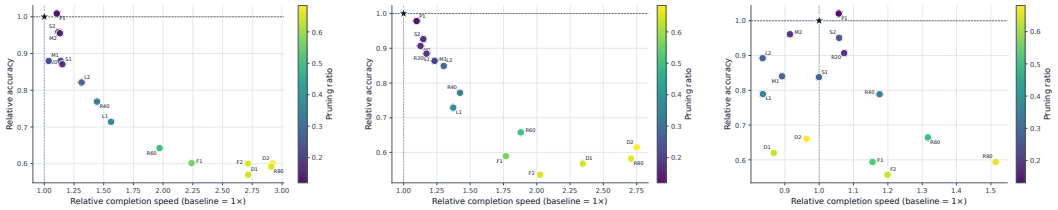


(a) Llama-3.2-3B-Instruct. (b) LLaMA-3.1-8B-Instruct. (c) Mistral-7B-Instruct-v0.3.

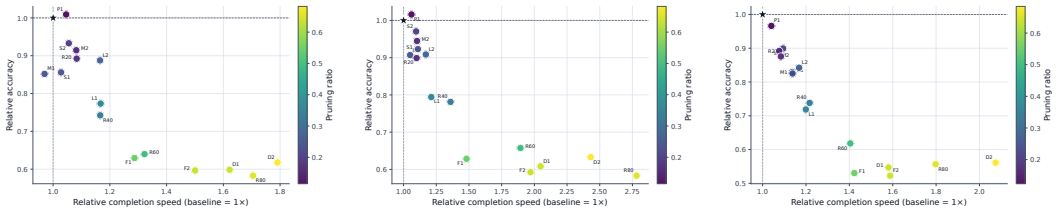
(d) Qwen2.5-7B-Instruct. (e) Qwen2.5-14B-Instruct. (f) Qwen3.5-4B.

(g) Qwen3.5-9B. (h) Phi-4. (i) Qwen3-4B-Thinking-2507.

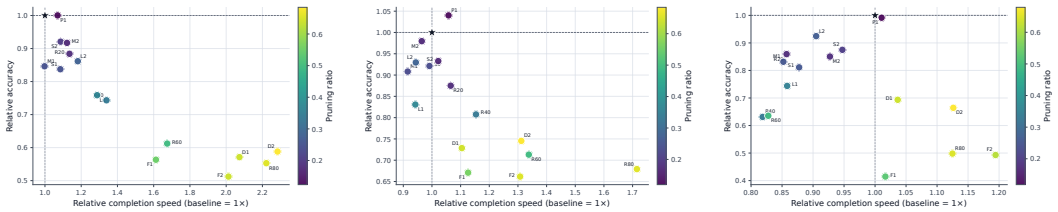
Figure 5: **Performance–efficiency landscape.** Relative task accuracy (Q_{rel}, macro mean over six benchmarks) over token retention ratio ($1 - \rho$); color and dashed iso-score hyperbolas (best, mean, worst) indicate $S_c = \overline{Q_{rel}}\,\overline{\rho}$, with both quantities first macro-averaged with equal weight across the six benchmarks.



(a) Llama-3.2-3B-Instruct. (b) LLaMA-3.1-8B-Instruct. (c) Mistral-7B-Instruct-v0.3.

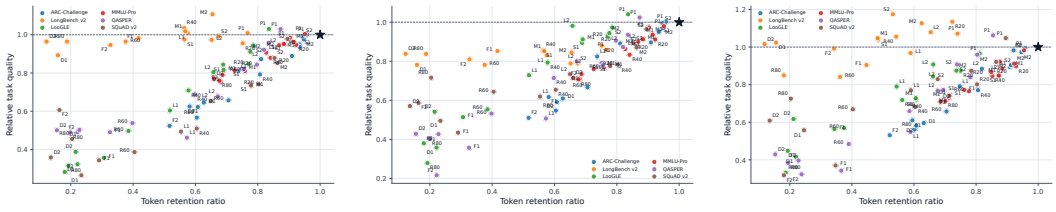


(d) Qwen2.5-7B-Instruct. (e) Qwen2.5-14B-Instruct. (f) Qwen3.5-4B.



(g) Qwen3.5-9B. (h) Phi-4. (i) Qwen3-4B-Thinking-2507.

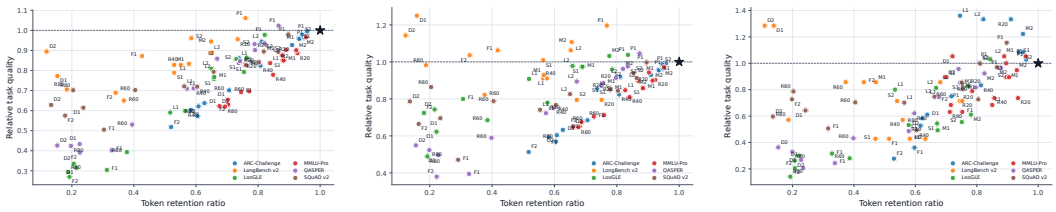
Figure 6: **End-to-end speed-up induced by static pruning.** Relative task accuracy (Q_{rel}) over relative task completion speed (baseline t_{TC} / pruned t_{TC}); color encodes the pruning ratio ρ.



(a) Llama-3.2-3B-Instruct. (b) LLaMA-3.1-8B-Instruct. (c) Mistral-7B-Instruct-v0.3.



(d) Qwen2.5-7B-Instruct. (e) Qwen2.5-14B-Instruct. (f) Qwen3.5-4B.



(g) Qwen3.5-9B. (h) Phi-4. (i) Qwen3-4B-Thinking-2507.

Figure 7: **Per-benchmark performance–efficiency landscape.** Relative task quality over token retention ratio $(1 - \rho)$, with color encoding the benchmark. Points above the dashed line outperform the unpruned baseline; most of them belong to LongBench-v2.

C.3 SENSITIVITY TO ACCURACY WEIGHTING

For benchmark b, let $Q_{\mathrm{rel},b}$ denote relative task quality and ρ_b the pruning ratio. We first compute unweighted macro means over the six benchmarks,

$$\bar{Q}_{\mathrm{rel}} = \frac{1}{6}\sum_{b=1}^{6} Q_{\mathrm{rel},b}, \qquad \bar{\rho} = \frac{1}{6}\sum_{b=1}^{6} \rho_b, \tag{1}$$

and define

$$S_{c,\mathrm{w}} = \bar{Q}_{\mathrm{rel}}^{\mathrm{w}}\bar{\rho}. \tag{2}$$

Thus, benchmarks receive equal weight regardless of their sample count. The exponent w controls the score's sensitivity to retained utility: in log space, $\log S_{c,\mathrm{w}} = \mathrm{w}\log\bar{Q}_{\mathrm{rel}} + \log\bar{\rho}$. At $\mathrm{w} = 1$, proportional changes in retained quality and reduction contribute equally, whereas larger w penalizes quality loss more strongly.

Table 15 reports the resulting rankings for Qwen3.5-9B. At $\mathrm{w} = 1$, the ranking is dominated by aggressive patterns (D2, D1, R80, and F2). At $\mathrm{w} = 2$, D2 remains first, L1 moves to second place, and L2 moves to fourth place ahead of R80, while both frequency rules fall outside the top eight. D2 is the only pattern ranked first under both settings, identifying it as the most robust aggressive reduction for the discovery model. Because $S_{c,\mathrm{w}}$ rewards reduction irrespective of how it is achieved, we interpret it together with random deletion at comparable pruning rates.

Table 15: **Pattern ranking for Qwen3.5-9B under different quality exponents** w**.** $S_{c,\mathrm{w}} = \bar{Q}_{\mathrm{rel}}^{\mathrm{w}}\bar{\rho}$, where both quantities are unweighted macro means over six benchmarks.

	w = 1				w = 2			
Rank	Pattern	Q_{rel}	ρ	$S_{c,1}$	Pattern	Q_{rel}	ρ	$S_{c,2}$
1	D2	0.588	0.684	0.402	D2	0.588	0.684	0.237
2	D1	0.571	0.640	0.366	L1	0.759	0.366	0.211
3	R80	0.553	0.655	0.362	D1	0.571	0.640	0.209
4	F2	0.512	0.640	0.328	L2	0.862	0.281	0.209
5	F1	0.563	0.550	0.310	R80	0.553	0.655	0.200
6	R60	0.612	0.496	0.304	R40	0.743	0.339	0.187
7	L1	0.759	0.366	0.278	R60	0.612	0.496	0.186
8	R40	0.743	0.338	0.252	M1	0.846	0.249	0.178
9	L2	0.862	0.281	0.242	S2	0.921	0.208	0.176
10	M1	0.846	0.249	0.211	S1	0.837	0.250	0.176
11	S1	0.837	0.250	0.210	F1	0.563	0.550	0.175
12	S2	0.921	0.208	0.191	F2	0.512	0.640	0.168
13	R20	0.884	0.180	0.159	M2	0.917	0.172	0.145
14	M2	0.917	0.172	0.158	R20	0.884	0.180	0.141
15	P1	1.000	0.124	0.124	P1	1.000	0.124	0.124

C.4 CROSS-MODEL RANKING STABILITY

To quantify transfer, we rank the interventions for each model and compute Spearman and Kendall correlations as well as the fraction of identically ordered intervention pairs for all 36 pairs of models (Tab. 16). Since a rule's reduction rate is nearly constant across models, rankings by ρ alone are almost identical ($\rho_s \geq 0.975$; Tab. 17), and the high agreement of S_c-based rankings is partly mechanical. Rankings by relative quality alone remain highly correlated (median $\rho_s = 0.957$), and rankings of the 11 linguistic rules by their residual to the linguistic trend are moderately correlated (median 0.777, minimum 0.300). At the benchmark level, relative-quality rankings are stable on QASPER, ARC-Challenge, and MMLU-Pro (median $\rho_s \geq 0.967$), but not on LongBench-v2 (median 0.308), where the minimum over model pairs is even negative (-0.563): on this benchmark the effect of an intervention is governed by how much of the input a model has to truncate, which differs by a factor of almost three across models (Tab. 31), so two models with very different context budgets can order the interventions in opposite ways.

C.5 STATISTICAL SIGNIFICANCE

Table 18 summarizes the paired tests. The large majority of significant differences are losses, as expected for input reduction. Significant improvements are rare and small; they concern P1 on

Table 16: **Cross-model stability of intervention rankings.** Statistics over all 36 pairs of the nine evaluated models (Spearman ρ_s, Kendall τ, and the fraction of intervention pairs ordered identically).

Scope	ρ_s min	ρ_s median	ρ_s max	τ median	Pair agreement (median)
Four context benchmarks (composite)	0.877	0.976	0.996	0.890	94.5%
All six benchmarks (composite)	0.894	0.963	0.996	0.868	93.4%
SQuAD v2	0.767	0.965	1.000	0.890	94.5%
LongBench-v2	0.648	0.910	0.987	0.791	89.6%
LooGLE	0.411	0.712	0.982	0.527	76.4%
QASPER	0.433	0.851	0.982	0.758	87.9%
MMLU-Pro	0.886	0.996	1.000	0.978	98.9%
ARC-C	0.723	0.991	1.000	0.956	97.8%

Table 17: **Cross-model ranking stability for different ranking criteria.** Spearman correlation of intervention rankings over all 36 pairs of the nine models. Rankings by the pruning ratio alone are almost identical across models, so part of the stability of S_c-based rankings is mechanical; rankings by relative quality and by reduction-adjusted residual quality isolate the behavioral component.

Ranking criterion (macro over six benchmarks)	min	median	max
Pruning ratio ρ (reference)	0.975	0.986	1.000
Relative quality Q_{rel}	0.786	0.957	0.993
Residual quality, linguistic trend (Table 13)	0.300	0.777	0.973
Relative quality Q_{rel} per benchmark			
SQuAD v2	0.721	0.905	0.989
LongBench-v2	−0.563	0.308	0.907
LooGLE	0.696	0.896	0.982
QASPER	0.857	0.973	0.996
MMLU-Pro	0.415	0.967	0.999
ARC-C	0.705	0.989	0.999

QASPER and SQuAD v2, three conservative rules for Qwen2.5-14B on LooGLE, and R80 for Phi-4 on the longest LongBench-v2 documents (Tab. 19). On LongBench-v2, 27 comparisons are significantly worse on short documents, whereas none is significantly worse on long documents, which supports the coverage account in Section 4.4.

Table 18: **Paired significance tests.** Number of model–intervention pairs whose paired per-sample difference to the unpruned baseline survives Benjamini–Hochberg correction ($q \leq 0.05$), split by sign. LongBench-v2 is additionally tested within its length categories.

Benchmark (scope)	Significantly worse	Significantly better
arc_challenge	111	0
longbench_v2	15	0
longbench_v2 :: long	0	1
longbench_v2 :: short	27	0
loogle_longdep_qa	109	3
mmlu_pro	118	0
qasper	122	4
squad_v2	127	1
xquad_de	48	0
xquad_en	46	0
xquad_tr	47	0

C.6 Reasoning model

The efficiency profile differs for the reasoning-tuned Qwen3-4B-Thinking-2507. Baseline completion times range from 31.6 s (SQuAD v2) to 255.7 s (LongBench-v2), compared with at most about 15 s for the non-reasoning models, because completion is dominated by the autoregressive generation of the reasoning trace. Pruning still accelerates the prefill stage considerably (e.g., $10.8\times$ faster t_{TTFT} for D2 on LongBench-v2), but the end-to-end speed-up is modest ($1.13\times$ for D2 on average) and nine of the 15 interventions are slower than the baseline on average (Tab. 12). Relative quality follows the same hierarchy as for the non-reasoning models (Spearman ρ_s between 0.88 and 0.95 with every other model on the four-benchmark composite), although the small subsets make individual estimates noisy. On LongBench-v2, D1 and D2 score 18% versus 14% for the baseline on the 50-item subset.

Table 19: **All static interventions that significantly improve over the unpruned baseline** after Benjamini–Hochberg correction ($q \leq 0.05$).

Model	Benchmark (scope)	Pattern	n	Δ	q
Phi-4	longbench_v2 :: long	R80	108	+0.1204	0.0351
Qwen2.5-14B	loogle_longdep_qa	S1	1101	+0.0089	0.0028
Qwen2.5-14B	loogle_longdep_qa	S2	1101	+0.0126	0.0004
Qwen2.5-14B	loogle_longdep_qa	M1	1101	+0.0103	0.0020
Qwen2.5-14B	qasper	P1	1451	+0.0128	0.0078
Qwen2.5-7B	qasper	P1	1451	+0.0141	0.0088
Phi-4	qasper	P1	1451	+0.0136	0.0045
Mistral-7B	qasper	P1	1451	+0.0198	0.0006
Mistral-7B	squad_v2	P1	11873	+0.0200	0.0001

Tables 20 and 21 show the beginning of the reasoning trace for the same LongBench-v2 item without and with D1 pruning. With the dense input, the model summarizes each document and evaluates the options against them. With the pruned input, it first recognizes that the context is "a mix of keywords, phrases, and references" and then reconstructs the topic from salient entities (e.g., the Xu case, Ginsburg, ARTs). This illustrates how a reasoning model can partially re-assemble the semantic nucleus from a telegraphic input, at the cost of additional reasoning tokens.

Table 20: **Exemplary reasoning trace of Qwen3-4B-Thinking-2507.** The trace is recorded for the unpruned baseline input on a LongBench-v2 item (truncated for space).

"I need to analyze the two documents to determine what information can be drawn from them. Let me first understand what each document is about.
Document 1: This is a legal analysis of China's regulatory framework on assisted reproductive technologies (ARTs). It discusses how China has prohibited unmarried women from accessing ARTs for over two decades, with specific references to the Ethical Principles for Human Reproductive Technologies and Human Sperm Banks. It mentions a case called Zaozao Xu's case where an unmarried woman challenged the prohibition on egg freezing. The document discusses how this prohibition constitutes discrimination against women, legislative conflicts in the regulatory framework, and the need for reform.
Document 2: This is an analysis of Ruth Bader Ginsburg's work on sex equality perspectives on reproductive rights. It discusses how Ginsburg viewed laws that deprive women of control over reproductive choices as raising questions of sex equality. It covers her legal career, her views on how reproductive rights relate to equal protection, and how the Supreme Court has responded to these equality arguments over time. It also discusses the concept of sex discrimination in the context of reproductive regulation.
Now I need to evaluate each option to see what information can be drawn from the two documents.
Option A: ""Discrimination against women can be seen in the Ethical Principles for Human Reproductive Technologies and Human Sperm Banks alone.""
From Document 1, we learn that Section 1 Article 4(1) of the Ethical Principles for Human Reproductive Technologies and Human Sperm Banks explicitly states that ""couples who do not comply with national population and family planning laws and regulations and single women"" are prohibited from accessing ARTs. The document states that this prohibition constitutes direct discrimination against women. However, the document also mentions that there are other factors contributing to discrimination against women in the regulatory framework, such as the Administrative Measures for ARTs and the conflict between different regulations. So while discrimination is indeed present in the Ethical Principles, it's not the only factor. The document says ""the current regulatory framework on ARTs has remained unchanged and unchallenged for over two decades, prohibiting unmarried women from accessing any form of ARTs"" and that ""unmarried men are allowed to freeze sperm regardless of their marital status."" So there's more to the discrimination than just the Ethical Principles. The document also mentions ""significant legislative conflicts"" and ""deeply rooted discrimination against women."" So this option seems too narrow - it's not just the Ethical Principles that show discrimination.
Option B: """

Table 21: **Exemplary reasoning trace of Qwen3-4B-Thinking-2507.** The trace is recorded for the same item after D1 pruning (truncated for space). Blue text highlights where the model explicitly recognizes the telegraphic input.

"First, I need to understand the context. It's a long string of text that seems to be a mix of keywords, phrases, and references related to women's rights, ARTs (Assisted Reproductive Technologies), Chinese regulations, legal cases, and Ruth Bader Ginsburg's work on gender equality and reproductive rights.
The question is: ""information documents"" and the options are:
A. Discrimination women Principles Technologies
B. Ginsberg women result nation
C. factors China women eggs
D. Ginsburg leave women
I need to figure out which one is the best answer based on the context provided.
Let me read through the context to see what it's about. The context starts with ""Analysis article technologies decades women form ARTs..."" and it's full of terms like ""ARTs"", ""China"", ""Xu case"", ""Court"", ""Ginsburg"", ""fertility"", ""egg freezing"", ""laws"", ""regulations"", etc.
It mentions ""Xu case"" which is a specific legal case in China about egg freezing and women's rights. There's a reference to ""Ginsburg"" which is Ruth Bader Ginsburg, the Supreme Court Justice known for her work on gender equality and reproductive rights.
Also, there's a lot about Chinese regulations on ARTs, the ""Xu case"" where a woman was refused services for egg freezing, and the legal battles around it.
The context has phrases like ""Ginsburg essay perspective"", ""1970s"", ""Roe v. Wade"", ""abortion"", ""reproductive rights"", and it talks about how Ginsburg's views on equality and rights relate to current issues.
Specifically, it mentions: ""This important quarter century Ginsburg essay perspective rights that Ginsburg decades lawyer career Ginsburg laws that women control they children questions equality Ginsburg 1970s organization work state women control decision them citizenship decades law Ginsburg she claims choice values reasoning reproduction law abortion discrimination role doctrines access contraception""
And later: ""Ginsburg leave women"" appears in the options.
Also, in the context, there's a part about ""FMLA"" (Family and Medical Leave Act) and how Ginsburg discussed discrimination against women in the workplace related to pregnancy and childcare.
Let me look for clues about what ""information documents"" refers to. The question says ""information documents"" but it's a bit vague. In the context, it might be referring to the content of the documents or the focus.

C.7 Absolute results per model

Tables 22–30 report absolute scores, warm-cache completion times t_{TC} (seconds), and pruning ratios ρ (%) for every model, benchmark, and intervention.

Table 22: **Pruning patterns on Llama-3.2-3B-Instruct.** Task metrics, warm-cache completion time (t_{TC}), and pruning ratio (ρ) are reported.

Category	Method	LongBench-v2			SQuAD v2			
		$Q_{abs}\uparrow$	$t_{TC}\downarrow$	ρ	EM ↑	F1 ↑	$t_{TC}\downarrow$	ρ
	Baseline	22.27	10.133	–	9.25	41.16	0.225	–
Structural Subword	S1	21.71	9.039	43.50	9.38	30.51	0.195	22.52
	S2	22.07	8.289	32.61	8.59	36.16	0.203	14.63
Punctuation	P1	22.47	8.316	23.25	8.68	40.27	0.210	10.47
Lexical & POS	L1	22.47	6.558	42.48	4.72	20.31	0.221	44.63
	L2	21.67	7.336	34.83	5.10	26.90	0.215	35.01
Frequency	F1	21.87	5.024	58.28	10.06	14.13	0.208	70.93
	F2	21.07	4.069	67.26	23.29	24.95	0.201	83.62
Syntactic	D1	19.88	1.878	84.12	6.04	10.91	0.172	76.59
	D2	21.47	1.507	87.76	11.46	14.75	0.119	86.38
Morphology	M1	22.66	15.174	43.27	9.24	30.22	0.196	22.14
	M2	24.65	7.293	34.53	9.58	35.07	0.215	13.67
Random Reference	R20	21.27	8.246	24.85	8.38	30.50	0.205	19.67
	R40	23.06	6.499	43.62	8.63	20.98	0.181	39.60
	R60	21.47	4.643	62.40	10.37	15.90	0.155	59.54
	R80	21.47	2.601	81.20	16.57	18.73	0.125	79.57

Category	Method	LooGLE			QASPER			
		ROUGE-L ↑	$t_{TC}\downarrow$	ρ	EM ↑	F1 ↑	$t_{TC}\downarrow$	ρ
	Baseline	11.24	3.009	–	4.48	33.71	0.755	–
Structural Subword	S1	9.49	2.091	31.12	4.27	27.64	0.646	24.90
	S2	10.61	2.362	21.45	3.10	32.28	0.668	16.46
Punctuation	P1	11.58	2.483	16.40	3.17	34.72	0.674	12.70
Lexical & POS	L1	6.79	1.701	48.18	7.79	15.58	0.247	42.74
	L2	9.05	2.010	34.22	3.24	22.84	0.383	32.84
Frequency	F1	4.01	1.002	69.35	14.61	16.52	0.140	67.41
	F2	3.18	0.651	81.91	16.20	16.53	0.124	77.80
Syntactic	D1	3.63	1.275	77.84	15.51	16.96	0.144	77.21
	D2	4.36	1.114	78.44	13.85	16.87	0.180	84.44
Morphology	M1	9.18	2.169	30.96	4.34	27.73	0.675	24.45
	M2	10.24	2.349	22.38	4.07	31.07	0.695	18.01
Random Reference	R20	9.53	2.289	22.93	5.24	28.45	0.598	20.28
	R40	7.97	1.707	42.23	10.06	23.19	0.361	40.23
	R60	5.59	1.248	61.48	14.06	18.14	0.200	60.20
	R80	3.55	0.888	80.75	16.26	16.51	0.119	80.11

Category	Method	MMLU-Pro			ARC-Challenge		
		$Q_{abs}\uparrow$	$t_{TC}\downarrow$	ρ	$Q_{abs}\uparrow$	$t_{TC}\downarrow$	ρ
	Baseline	28.19	0.057	–	69.11	0.057	–
Structural Subword	S1	26.82	0.060	11.64	65.61	0.058	6.60
	S2	27.76	0.059	9.67	68.86	0.057	5.35
Punctuation	P1	28.30	0.057	4.88	69.20	0.057	4.19
Lexical & POS	L1	25.56	0.057	17.30	55.89	0.058	26.29
	L2	26.65	0.057	13.12	60.32	0.058	18.93
Frequency	F1	22.91	0.058	27.41	43.17	0.058	41.87
	F2	21.20	0.058	31.85	36.18	0.057	48.30
Syntactic	D1	22.27	0.057	31.32	44.62	0.057	37.32
	D2	21.75	0.057	34.06	43.00	0.057	39.25
Morphology	M1	26.84	0.062	11.62	64.68	0.058	6.52
	M2	27.14	0.057	8.52	67.49	0.058	5.29
Random Reference	R20	26.70	0.057	7.65	61.43	0.057	9.05
	R40	24.79	0.057	15.91	54.78	0.057	19.29
	R60	22.86	0.057	24.08	45.39	0.057	29.40
	R80	21.42	0.056	32.31	39.16	0.056	39.56

Table 23: **Pruning patterns on Llama-3.1-8B-Instruct.** Task metrics, warm-cache completion time (t_{TC}), and pruning ratio (ρ) are reported.

Category	Method	LongBench-v2			SQuAD v2			
		$Q_{abs}\uparrow$	$t_{TC}\downarrow$	ρ	EM ↑	F1 ↑	$t_{TC}\downarrow$	ρ
	Baseline	28.23	15.154	–	29.12	62.13	0.209	–
Structural Subword	S1	24.10	9.845	43.50	26.24	48.48	0.188	22.52
	S2	22.27	11.506	32.61	26.43	55.42	0.197	14.63
Punctuation	P1	24.25	12.514	23.25	26.99	59.94	0.204	10.47
Lexical & POS	L1	23.46	9.932	42.48	19.35	38.49	0.168	44.63
	L2	22.27	11.096	34.83	21.28	45.60	0.164	35.01
Frequency	F1	24.06	7.626	58.28	22.73	27.00	0.156	70.93
	F2	22.86	6.171	67.26	35.04	36.46	0.154	83.62
Syntactic	D1	22.07	2.949	84.12	26.93	30.76	0.133	76.59
	D2	23.66	2.377	87.76	33.30	35.45	0.123	86.38
Morphology	M1	24.06	9.932	43.27	25.88	48.18	0.188	22.14
	M2	23.86	11.149	34.53	26.51	54.09	0.199	13.67
Random Reference	R20	24.85	12.463	24.85	25.44	48.66	0.189	19.67
	R40	24.25	9.947	43.62	28.77	40.64	0.156	39.60
	R60	22.07	7.206	62.40	36.27	40.02	0.132	59.54
	R80	23.66	4.075	81.20	43.60	44.48	0.122	79.57

Category	Method	LooGLE			QASPER			
		ROUGE-L ↑	$t_{TC}\downarrow$	ρ	EM ↑	F1 ↑	$t_{TC}\downarrow$	ρ
	Baseline	14.65	4.217	–	4.62	38.30	1.072	–
Structural Subword	S1	13.10	2.778	31.12	4.34	30.29	0.873	24.90
	S2	14.26	3.210	21.45	4.82	35.84	0.917	16.46
Punctuation	P1	15.25	3.442	16.40	4.96	39.22	0.965	12.70
Lexical & POS	L1	10.68	2.184	48.18	3.03	19.42	0.702	42.74
	L2	14.38	2.671	34.22	3.38	30.60	0.676	32.84
Frequency	F1	7.53	1.493	69.35	7.65	13.69	0.430	67.41
	F2	5.57	0.999	81.91	5.44	8.28	0.504	77.80
Syntactic	D1	5.25	1.524	77.84	11.23	16.34	0.414	77.21
	D2	7.93	1.301	78.44	9.65	16.41	0.338	84.44
Morphology	M1	13.36	2.796	30.96	4.27	29.84	0.877	24.45
	M2	13.84	3.290	22.38	3.93	33.44	1.013	18.01
Random Reference	R20	13.63	3.062	22.93	4.96	34.69	0.812	20.28
	R40	11.63	2.292	42.23	5.72	27.34	0.591	40.23
	R60	8.11	1.602	61.48	10.13	20.39	0.363	60.20
	R80	4.09	1.064	80.75	12.96	15.47	0.222	80.11

Category	Method	MMLU-Pro			ARC-Challenge		
		$Q_{abs}\uparrow$	$t_{TC}\downarrow$	ρ	$Q_{abs}\uparrow$	$t_{TC}\downarrow$	ρ
	Baseline	36.18	0.065	–	78.24	0.063	–
Structural Subword	S1	33.01	0.065	11.64	74.49	0.063	6.60
	S2	35.21	0.064	9.67	77.82	0.063	5.35
Punctuation	P1	35.44	0.065	4.88	78.41	0.063	4.19
Lexical & POS	L1	31.22	0.065	17.30	64.51	0.063	26.29
	L2	32.58	0.065	13.12	69.71	0.063	18.93
Frequency	F1	27.49	0.066	27.41	48.29	0.065	41.87
	F2	25.69	0.065	31.85	39.93	0.064	48.30
Syntactic	D1	26.57	0.064	31.32	47.70	0.063	37.32
	D2	25.90	0.064	34.06	46.25	0.063	39.25
Morphology	M1	33.02	0.064	11.62	74.40	0.063	6.52
	M2	33.88	0.065	8.52	75.94	0.063	5.29
Random Reference	R20	32.97	0.064	7.65	70.14	0.063	9.05
	R40	30.14	0.064	15.91	60.67	0.063	19.29
	R60	27.78	0.064	24.08	52.13	0.063	29.40
	R80	25.58	0.064	32.31	42.83	0.063	39.56

Table 24: **Pruning patterns on Mistral-7B-Instruct-v0.3.** Task metrics, warm-cache completion time (t_{TC}), and pruning ratio (ρ) are reported.

Category	Method	LongBench-v2			SQuAD v2			
		$Q_{abs}\uparrow$	$t_{TC}\downarrow$	ρ	EM ↑	F1 ↑	$t_{TC}\downarrow$	ρ
	Baseline	25.05	3.773	–	18.28	41.82	0.577	–
Structural Subword	S1	26.04	5.625	51.46	16.43	31.14	0.551	28.43
	S2	29.42	3.871	46.82	17.71	36.54	0.557	21.57
Punctuation	P1	26.84	3.441	25.97	19.63	43.82	0.557	10.45
Lexical & POS	L1	24.25	17.155	41.11	19.91	32.21	0.587	41.00
	L2	27.04	17.297	34.62	18.53	34.68	0.576	32.36
Frequency	F1	22.66	2.885	55.29	11.23	15.48	0.590	65.26
	F2	24.85	2.456	65.75	11.03	13.29	0.643	81.88
Syntactic	D1	25.65	15.656	84.42	19.73	23.31	0.559	75.41
	D2	25.45	15.410	88.09	23.20	25.46	0.491	86.46
Morphology	M1	26.24	12.436	51.71	16.38	30.98	0.556	28.61
	M2	28.23	8.659	37.51	19.22	38.52	0.584	13.53
Random Reference	R20	28.43	3.247	27.65	17.95	33.52	0.588	19.69
	R40	26.44	2.870	45.71	19.73	28.48	0.586	39.61
	R60	21.07	2.449	63.80	23.75	27.98	0.561	59.69
	R80	21.27	1.792	81.89	28.46	30.35	0.514	79.66

Category	Method	LooGLE			QASPER			
		ROUGE-L ↑	$t_{TC}\downarrow$	ρ	EM ↑	F1 ↑	$t_{TC}\downarrow$	ρ
	Baseline	14.83	4.032	–	3.38	31.87	1.338	–
Structural Subword	S1	10.69	3.474	39.30	3.51	24.51	1.192	30.50
	S2	12.51	3.305	33.86	4.41	28.92	1.186	26.14
Punctuation	P1	14.32	3.727	18.86	4.62	33.85	1.181	14.54
Lexical & POS	L1	11.72	4.999	45.59	3.38	17.47	1.267	41.12
	L2	13.46	5.312	33.78	3.93	24.49	1.228	32.38
Frequency	F1	8.36	2.711	65.61	1.52	10.93	1.197	63.41
	F2	6.65	2.541	80.64	4.27	10.29	1.202	76.21
Syntactic	D1	6.15	4.164	77.96	5.86	12.63	1.346	77.19
	D2	9.15	3.804	78.78	5.10	13.67	1.040	84.67
Morphology	M1	10.80	3.984	39.44	3.72	24.61	1.318	30.56
	M2	12.97	3.955	26.35	4.69	30.58	1.254	19.64
Random Reference	R20	12.96	3.611	25.00	3.03	26.76	1.162	21.88
	R40	10.65	2.943	43.77	4.00	21.00	1.002	41.44
	R60	8.44	2.218	62.50	4.14	15.44	0.937	61.01
	R80	4.92	1.753	81.28	5.24	12.23	0.929	80.51

Category	Method	MMLU-Pro			ARC-Challenge		
		$Q_{abs}\uparrow$	$t_{TC}\downarrow$	ρ	$Q_{abs}\uparrow$	$t_{TC}\downarrow$	ρ
	Baseline	29.75	0.197	–	73.29	0.198	–
Structural Subword	S1	25.28	0.198	12.55	66.21	0.198	8.68
	S2	27.45	0.197	12.40	72.01	0.200	7.92
Punctuation	P1	29.24	0.197	4.55	72.87	0.199	4.15
Lexical & POS	L1	25.80	0.202	15.11	58.02	0.203	25.25
	L2	26.33	0.204	11.54	64.85	0.203	18.13
Frequency	F1	22.98	0.192	23.80	44.71	0.196	40.51
	F2	21.73	0.188	28.72	38.91	0.195	47.80
Syntactic	D1	21.76	0.200	28.74	43.69	0.201	36.86
	D2	21.12	0.195	31.46	42.66	0.197	39.25
Morphology	M1	25.24	0.199	12.70	66.30	0.199	8.75
	M2	27.14	0.199	7.70	71.33	0.200	5.39
Random Reference	R20	26.71	0.193	7.13	65.53	0.197	9.14
	R40	25.19	0.192	14.76	56.48	0.194	19.39
	R60	22.76	0.187	22.28	48.21	0.192	29.53
	R80	21.21	0.183	29.87	41.13	0.192	39.79

Table 25: **Pruning patterns on Qwen2.5-7B-Instruct.** Task metrics, warm-cache completion time (t_{TC}), and pruning ratio (ρ) are reported.

Category	Method	LongBench-v2			SQuAD v2			
		$Q_{abs}\uparrow$	$t_{TC}\downarrow$	ρ	EM ↑	F1 ↑	$t_{TC}\downarrow$	ρ
	Baseline	30.22	4.156	–	38.01	72.21	0.522	–
Structural Subword	S1	26.64	5.590	45.75	33.97	56.52	0.485	24.32
	S2	31.01	4.000	39.12	34.91	62.99	0.503	18.18
Punctuation	P1	31.81	3.671	22.00	36.64	70.78	0.505	10.20
Lexical & POS	L1	26.44	3.381	40.56	29.04	49.04	0.520	43.57
	L2	29.22	3.516	33.19	30.92	56.45	0.488	34.16
Frequency	F1	30.22	2.894	55.91	26.37	31.68	0.549	69.34
	F2	28.23	2.544	64.64	36.58	38.34	0.458	81.88
Syntactic	D1	27.83	1.670	84.25	38.03	40.96	0.386	76.19
	D2	26.84	1.433	87.84	40.77	42.36	0.357	86.65
Morphology	M1	28.03	11.713	45.69	33.88	56.31	0.476	23.95
	M2	27.83	3.355	33.51	34.43	62.49	0.506	13.38
Random Reference	R20	30.02	3.611	24.59	34.69	57.33	0.504	19.68
	R40	25.65	3.215	43.42	34.48	46.09	0.472	39.61
	R60	24.85	2.736	62.27	39.24	43.19	0.411	59.56
	R80	27.24	1.940	81.13	46.28	47.02	0.352	79.58

Category	Method	LooGLE			QASPER			
		ROUGE-L ↑	$t_{TC}\downarrow$	ρ	EM ↑	F1 ↑	$t_{TC}\downarrow$	ρ
	Baseline	10.81	8.395	–	8.68	43.21	1.960	–
Structural Subword	S1	8.70	7.305	32.99	8.27	35.40	1.497	26.25
	S2	9.36	7.632	25.33	8.27	38.97	1.607	19.37
Punctuation	P1	10.92	8.138	16.00	8.06	44.62	1.768	12.47
Lexical & POS	L1	9.12	7.022	46.91	7.51	25.38	1.289	41.97
	L2	10.47	7.575	33.31	7.93	35.35	1.241	32.26
Frequency	F1	7.65	6.171	67.67	6.69	15.40	0.981	66.24
	F2	6.61	5.756	80.08	12.06	15.31	0.717	76.48
Syntactic	D1	4.96	6.018	78.01	12.13	17.02	0.767	77.15
	D2	6.63	5.242	78.80	10.54	18.12	0.696	84.52
Morphology	M1	8.32	7.374	32.84	8.06	34.66	1.527	25.81
	M2	9.66	8.000	21.96	7.99	40.13	1.697	17.85
Random Reference	R20	9.63	7.973	22.87	7.99	36.70	1.589	20.26
	R40	7.86	7.148	42.19	7.86	27.43	1.340	40.23
	R60	6.04	6.157	61.44	10.13	21.10	1.092	60.20
	R80	4.19	4.807	80.74	12.68	16.99	0.696	80.11

Category	Method	MMLU-Pro			ARC-Challenge		
		$Q_{abs}\uparrow$	$t_{TC}\downarrow$	ρ	$Q_{abs}\uparrow$	$t_{TC}\downarrow$	ρ
	Baseline	41.79	0.183	–	89.42	0.174	–
Structural Subword	S1	37.17	0.194	11.76	85.67	0.185	6.78
	S2	39.72	0.189	10.94	87.71	0.181	5.77
Punctuation	P1	40.91	0.181	4.46	89.42	0.183	4.15
Lexical & POS	L1	34.43	0.177	16.00	74.40	0.171	26.07
	L2	36.81	0.177	12.12	81.31	0.171	18.77
Frequency	F1	28.96	0.180	25.38	51.96	0.182	41.53
	F2	26.75	0.179	29.54	45.48	0.172	47.91
Syntactic	D1	26.99	0.185	29.70	54.01	0.183	37.21
	D2	26.32	0.180	32.38	50.94	0.184	39.15
Morphology	M1	36.79	0.184	11.74	85.15	0.183	6.70
	M2	38.51	0.180	7.62	85.58	0.174	5.22
Random Reference	R20	37.72	0.180	7.25	82.68	0.173	9.01
	R40	33.72	0.182	15.09	71.42	0.182	19.21
	R60	29.73	0.179	22.83	59.04	0.181	29.29
	R80	25.93	0.179	30.61	48.55	0.172	39.43

Table 26: **Pruning patterns on Qwen2.5-14B-Instruct.** Task metrics, warm-cache completion time (t_{TC}), and pruning ratio (ρ) are reported.

Category	Method	LongBench-v2			SQuAD v2			
		$Q_{abs}\uparrow$	$t_{TC}\downarrow$	ρ	EM ↑	F1 ↑	$t_{TC}\downarrow$	ρ
	Baseline	34.79	6.114	–	41.79	74.39	1.689	–
Structural Subword	S1	33.00	6.147	45.75	39.84	60.21	1.484	24.32
	S2	33.20	5.373	39.12	39.89	65.97	1.543	18.18
Punctuation	P1	37.38	5.516	22.00	41.36	73.59	1.559	10.20
Lexical & POS	L1	28.63	4.990	40.56	34.14	53.72	1.462	43.57
	L2	31.61	5.217	33.19	36.89	61.37	1.366	34.16
Frequency	F1	28.43	4.239	55.91	34.00	38.22	1.129	69.34
	F2	29.22	3.733	64.64	40.82	42.31	0.921	81.88
Syntactic	D1	28.23	2.502	84.25	43.70	45.48	0.740	76.19
	D2	29.62	2.124	87.84	46.52	47.29	0.629	86.65
Morphology	M1	30.22	9.706	45.69	39.71	60.04	1.478	23.95
	M2	31.01	5.078	33.51	40.82	67.53	1.484	13.38
Random Reference	R20	33.20	5.383	24.59	41.37	60.28	1.458	19.68
	R40	29.22	4.795	43.42	44.48	51.24	0.921	39.61
	R60	26.24	4.084	62.27	47.71	49.11	0.665	59.56
	R80	28.03	2.906	81.13	49.31	49.44	0.600	79.58

Category	Method	LooGLE			QASPER			
		ROUGE-L ↑	$t_{TC}\downarrow$	ρ	EM ↑	F1 ↑	$t_{TC}\downarrow$	ρ
	Baseline	9.91	7.094	–	6.00	35.19	2.057	–
Structural Subword	S1	10.80	5.245	32.99	7.65	30.83	1.772	26.25
	S2	11.16	5.851	25.33	7.99	34.00	1.850	19.37
Punctuation	P1	10.22	6.544	16.00	7.44	36.47	1.917	12.47
Lexical & POS	L1	8.58	4.778	46.91	5.44	24.04	1.470	41.97
	L2	10.17	5.484	33.31	7.65	31.36	1.568	32.26
Frequency	F1	6.97	3.718	67.67	9.44	16.83	0.994	66.24
	F2	5.45	3.177	80.08	14.33	16.40	0.498	76.48
Syntactic	D1	4.92	3.161	78.01	13.09	16.80	0.622	77.15
	D2	5.97	2.872	78.80	13.03	17.96	0.453	84.52
Morphology	M1	10.94	5.266	32.84	7.51	30.11	1.774	25.81
	M2	10.38	6.118	21.96	7.17	33.41	1.865	17.85
Random Reference	R20	8.90	6.307	22.87	7.65	32.55	1.765	20.26
	R40	8.22	5.168	42.19	10.13	25.70	1.243	40.23
	R60	6.26	3.820	61.44	13.92	18.99	0.596	60.20
	R80	3.85	2.707	80.74	16.06	16.85	0.289	80.11

Category	Method	MMLU-Pro			ARC-Challenge		
		$Q_{abs}\uparrow$	$t_{TC}\downarrow$	ρ	$Q_{abs}\uparrow$	$t_{TC}\downarrow$	ρ
	Baseline	50.71	0.098	–	93.34	0.095	–
Structural Subword	S1	43.54	0.097	11.76	89.25	0.095	6.78
	S2	46.09	0.097	10.94	91.47	0.095	5.77
Punctuation	P1	49.11	0.097	4.46	93.00	0.095	4.15
Lexical & POS	L1	42.08	0.097	16.00	78.58	0.095	26.07
	L2	44.61	0.097	12.12	85.92	0.095	18.77
Frequency	F1	34.56	0.098	25.38	53.84	0.098	41.53
	F2	31.33	0.097	29.54	47.87	0.096	47.91
Syntactic	D1	32.31	0.096	29.70	57.42	0.096	37.21
	D2	30.73	0.096	32.38	55.20	0.096	39.15
Morphology	M1	43.12	0.096	11.74	89.25	0.094	6.70
	M2	45.43	0.097	7.62	91.13	0.095	5.22
Random Reference	R20	45.20	0.097	7.25	85.32	0.095	9.01
	R40	39.95	0.096	15.09	75.68	0.095	19.21
	R60	34.77	0.096	22.83	62.88	0.095	29.29
	R80	30.24	0.096	30.61	52.82	0.095	39.43

Table 27: **Pruning patterns on Qwen3.5-4B.** Task metrics, warm-cache completion time (t_{TC}), and pruning ratio (ρ) are reported.

Category	Method	LongBench-v2			SQuAD v2			
		$Q_{abs}\uparrow$	$t_{TC}\downarrow$	ρ	EM ↑	F1 ↑	$t_{TC}\downarrow$	ρ
	Baseline	35.39	4.238	–	27.36	62.94	0.274	–
Structural Subword	S1	30.62	3.137	47.02	24.75	48.90	0.234	24.17
	S2	31.41	3.328	41.50	25.00	54.33	0.250	18.16
Punctuation	P1	36.18	3.692	24.04	24.63	60.00	0.265	10.25
Lexical & POS	L1	28.63	3.402	42.28	19.41	42.18	0.240	43.77
	L2	31.61	3.536	35.03	18.95	45.80	0.228	34.32
Frequency	F1	27.24	2.918	57.37	21.96	27.54	0.288	69.65
	F2	29.03	2.566	65.90	33.88	36.04	0.243	82.17
Syntactic	D1	28.83	1.641	84.65	28.59	32.78	0.200	76.29
	D2	28.23	1.389	88.16	35.81	37.94	0.173	86.68
Morphology	M1	30.62	3.149	47.03	24.53	48.56	0.233	23.84
	M2	32.80	3.375	35.13	23.77	53.03	0.251	13.48
Random Reference	R20	33.60	3.641	26.55	26.45	51.66	0.248	19.68
	R40	30.82	3.217	44.89	29.64	43.04	0.216	39.60
	R60	27.63	2.717	63.25	37.70	41.91	0.189	59.58
	R80	28.23	1.910	81.62	45.58	46.33	0.173	79.57

Category	Method	LooGLE			QASPER			
		ROUGE-L ↑	$t_{TC}\downarrow$	ρ	EM ↑	F1 ↑	$t_{TC}\downarrow$	ρ
	Baseline	11.00	5.772	–	4.34	44.79	1.266	–
Structural Subword	S1	8.26	5.066	34.20	3.65	35.22	1.114	26.24
	S2	9.92	5.197	27.00	3.51	38.44	1.195	21.12
Punctuation	P1	9.96	5.550	17.87	3.51	43.01	1.261	13.38
Lexical & POS	L1	7.08	4.665	48.31	2.14	24.74	0.817	42.90
	L2	9.97	5.085	35.01	2.41	32.71	0.870	33.18
Frequency	F1	4.57	4.105	68.73	2.48	13.48	0.465	67.18
	F2	4.01	3.698	80.78	8.96	13.45	0.398	77.42
Syntactic	D1	4.03	3.750	78.55	6.89	17.02	0.648	77.50
	D2	4.44	3.593	79.43	6.75	17.37	0.305	84.70
Morphology	M1	8.13	5.081	34.11	3.51	34.90	1.112	25.82
	M2	9.41	5.405	23.68	3.65	38.94	1.232	17.94
Random Reference	R20	9.58	5.275	24.57	3.58	36.67	1.125	20.98
	R40	7.15	4.661	43.46	4.34	29.21	0.875	40.77
	R60	5.49	4.040	62.28	4.96	18.83	0.650	60.56
	R80	3.55	3.445	81.16	9.92	16.14	0.388	80.28

Category	Method	MMLU-Pro			ARC-Challenge		
		$Q_{abs}\uparrow$	$t_{TC}\downarrow$	ρ	$Q_{abs}\uparrow$	$t_{TC}\downarrow$	ρ
	Baseline	46.73	0.106	–	92.24	0.098	–
Structural Subword	S1	39.59	0.104	11.80	87.63	0.097	6.77
	S2	42.33	0.104	11.00	90.19	0.097	5.77
Punctuation	P1	44.84	0.105	4.52	91.81	0.097	4.16
Lexical & POS	L1	38.32	0.104	16.07	75.26	0.097	26.13
	L2	40.94	0.105	12.19	84.81	0.097	18.82
Frequency	F1	31.44	0.105	25.48	54.01	0.098	41.64
	F2	28.36	0.105	29.62	43.43	0.098	48.00
Syntactic	D1	29.31	0.103	29.75	52.82	0.096	37.27
	D2	28.21	0.103	32.43	52.65	0.096	39.21
Morphology	M1	39.59	0.104	11.79	87.37	0.097	6.71
	M2	41.76	0.105	7.69	88.82	0.097	5.23
Random Reference	R20	41.45	0.105	7.31	83.53	0.097	9.02
	R40	35.86	0.104	15.14	74.15	0.097	19.23
	R60	31.51	0.103	22.88	61.60	0.096	29.32
	R80	27.83	0.103	30.65	48.46	0.096	39.49

Table 28: **Pruning patterns on Qwen3.5-9B.** Task metrics, warm-cache completion time (t_{TC}), and pruning ratio (ρ) are reported.

Category	Method	LongBench-v2			SQuAD v2			
		$Q_{abs}\uparrow$	$t_{TC}\downarrow$	ρ	EM ↑	F1 ↑	$t_{TC}\downarrow$	ρ
	Baseline	35.79	4.853	–	35.67	70.04	0.286	–
Structural Subword	S1	28.23	5.595	47.02	35.99	58.03	0.236	24.17
	S2	34.39	4.534	41.50	34.87	62.64	0.257	18.16
Punctuation	P1	37.97	4.275	24.04	34.05	68.63	0.271	10.25
Lexical & POS	L1	29.82	3.958	42.28	28.15	50.48	0.218	43.77
	L2	31.81	4.100	35.03	28.32	55.57	0.252	34.32
Frequency	F1	31.21	3.376	57.37	30.19	35.40	0.229	69.65
	F2	24.65	2.965	65.90	38.35	40.28	0.207	82.17
Syntactic	D1	27.63	1.956	84.65	40.33	43.03	0.172	76.29
	D2	32.01	1.657	88.16	42.76	44.00	0.169	86.68
Morphology	M1	29.62	11.566	47.03	36.08	57.90	0.237	23.84
	M2	33.80	3.919	35.13	34.05	62.52	0.250	13.48
Random Reference	R20	34.19	4.219	26.55	38.47	58.92	0.235	19.68
	R40	29.62	3.736	44.89	43.36	50.39	0.191	39.60
	R60	23.26	3.191	63.25	47.70	49.19	0.174	59.58
	R80	25.25	2.272	81.62	48.90	49.14	0.169	79.57

Category	Method	LooGLE			QASPER			
		ROUGE-L ↑	$t_{TC}\downarrow$	ρ	EM ↑	F1 ↑	$t_{TC}\downarrow$	ρ
	Baseline	12.09	6.307	–	6.34	41.39	1.584	–
Structural Subword	S1	9.16	5.510	34.20	6.48	34.65	1.214	26.24
	S2	10.36	5.716	27.00	6.69	38.55	1.272	21.12
Punctuation	P1	11.81	6.019	17.87	5.79	42.37	1.360	13.38
Lexical & POS	L1	7.14	5.019	48.31	5.17	29.38	0.828	42.90
	L2	9.78	5.486	35.01	5.58	35.54	0.980	33.18
Frequency	F1	3.68	4.426	68.73	8.55	16.64	0.446	67.18
	F2	3.27	3.949	80.78	13.92	16.20	0.291	77.42
Syntactic	D1	3.81	3.951	78.55	10.96	17.95	0.339	77.50
	D2	4.06	3.951	79.43	10.54	17.65	0.291	84.70
Morphology	M1	9.28	5.580	34.11	7.03	35.07	1.253	25.82
	M2	10.45	5.823	23.68	5.58	38.72	1.267	17.94
Random Reference	R20	9.58	5.768	24.57	6.20	37.32	1.183	20.98
	R40	7.24	5.094	43.46	8.34	29.43	0.802	40.77
	R60	4.75	4.432	62.28	11.23	21.95	0.463	60.56
	R80	3.55	3.761	81.16	14.61	17.56	0.274	80.28

Category	Method	MMLU-Pro			ARC-Challenge		
		$Q_{abs}\uparrow$	$t_{TC}\downarrow$	ρ	$Q_{abs}\uparrow$	$t_{TC}\downarrow$	ρ
	Baseline	51.97	0.107	–	94.54	0.099	–
Structural Subword	S1	44.17	0.109	11.80	90.87	0.099	6.77
	S2	46.91	0.109	11.00	92.49	0.098	5.77
Punctuation	P1	50.19	0.106	4.52	94.03	0.098	4.16
Lexical & POS	L1	43.48	0.105	16.07	81.40	0.098	26.13
	L2	45.50	0.106	12.19	89.25	0.100	18.82
Frequency	F1	35.99	0.106	25.48	57.08	0.099	41.64
	F2	32.74	0.105	29.62	48.98	0.099	48.00
Syntactic	D1	34.01	0.104	29.75	60.24	0.098	37.27
	D2	32.36	0.104	32.43	58.70	0.098	39.21
Morphology	M1	44.18	0.110	11.79	90.53	0.100	6.71
	M2	46.83	0.106	7.69	91.21	0.099	5.23
Random Reference	R20	46.06	0.106	7.31	87.63	0.098	9.02
	R40	40.48	0.105	15.14	77.90	0.098	19.23
	R60	36.18	0.105	22.88	66.30	0.098	29.32
	R80	32.22	0.104	30.65	54.10	0.097	39.49

Table 29: **Pruning patterns on Phi-4.** Task metrics, warm-cache completion time (t_{TC}), and pruning ratio (ρ) are reported.

Category	Method	LongBench-v2			SQuAD v2			
		$Q_{abs}\uparrow$	$t_{TC}\downarrow$	ρ	EM ↑	F1 ↑	$t_{TC}\downarrow$	ρ
	Baseline	22.27	3.354	–	32.49	54.40	0.562	–
Structural Subword	S1	22.47	5.408	43.62	30.01	46.72	0.498	22.55
	S2	17.69	3.539	32.84	30.00	51.08	0.533	14.67
Punctuation	P1	26.64	2.987	23.20	31.26	54.22	0.541	10.47
Lexical & POS	L1	20.28	9.796	42.39	24.27	40.31	0.592	44.61
	L2	23.66	9.893	34.76	25.81	44.93	0.566	34.99
Frequency	F1	23.66	2.534	58.22	20.94	25.62	0.618	70.89
	F2	23.06	2.234	67.19	34.56	36.14	0.397	83.56
Syntactic	D1	27.83	8.591	84.12	34.98	37.82	0.417	76.59
	D2	25.45	8.416	87.75	41.43	42.77	0.270	86.38
Morphology	M1	20.68	12.109	43.41	29.71	46.29	0.505	22.16
	M2	24.65	5.286	34.47	29.91	49.64	0.545	13.63
Random Reference	R20	17.69	2.911	24.87	29.98	45.90	0.522	19.66
	R40	20.28	2.594	43.63	33.00	41.75	0.454	39.60
	R60	18.29	2.242	62.41	39.91	42.85	0.328	59.54
	R80	21.87	1.769	81.20	46.42	46.98	0.220	79.57

Category	Method	LooGLE			QASPER			
		ROUGE-L ↑	$t_{TC}\downarrow$	ρ	EM ↑	F1 ↑	$t_{TC}\downarrow$	ρ
	Baseline	7.32	5.208	–	8.41	29.73	2.147	–
Structural Subword	S1	7.12	4.926	31.17	9.17	25.88	1.884	24.92
	S2	7.02	5.044	21.49	8.41	29.00	1.954	16.49
Punctuation	P1	7.60	5.003	16.42	8.61	31.08	1.944	12.70
Lexical & POS	L1	6.65	4.678	48.15	4.76	21.48	1.947	42.73
	L2	7.16	4.765	34.20	6.82	26.57	1.784	32.84
Frequency	F1	5.85	4.123	69.31	2.27	11.73	1.909	67.39
	F2	5.30	3.590	81.86	5.51	11.28	1.723	77.78
Syntactic	D1	4.56	3.568	77.84	6.13	14.73	1.678	77.21
	D2	5.44	3.568	78.44	9.24	16.31	1.222	84.44
Morphology	M1	7.13	4.970	31.01	9.30	26.26	1.934	24.47
	M2	7.56	4.903	22.38	8.68	28.61	1.917	18.02
Random Reference	R20	6.34	5.055	22.95	7.86	27.16	1.906	20.28
	R40	5.71	4.796	42.24	7.44	22.41	1.716	40.23
	R60	5.01	4.281	61.49	7.99	17.52	1.458	60.20
	R80	3.58	3.305	80.75	10.82	15.55	1.063	80.11

Category	Method	MMLU-Pro			ARC-Challenge		
		$Q_{abs}\uparrow$	$t_{TC}\downarrow$	ρ	$Q_{abs}\uparrow$	$t_{TC}\downarrow$	ρ
	Baseline	53.07	0.224	–	94.03	0.199	–
Structural Subword	S1	45.61	0.223	11.65	90.02	0.200	6.60
	S2	50.02	0.226	9.69	92.75	0.198	5.35
Punctuation	P1	51.53	0.220	4.88	93.34	0.195	4.19
Lexical & POS	L1	44.95	0.217	17.29	80.38	0.179	26.28
	L2	47.63	0.222	13.11	86.52	0.194	18.93
Frequency	F1	37.39	0.211	27.39	55.72	0.185	41.86
	F2	34.67	0.206	31.83	48.21	0.174	48.29
Syntactic	D1	35.82	0.209	31.31	59.47	0.185	37.31
	D2	34.42	0.206	34.05	56.74	0.183	39.24
Morphology	M1	45.50	0.223	11.63	89.76	0.211	6.52
	M2	47.69	0.228	8.53	90.70	0.208	5.28
Random Reference	R20	47.89	0.221	7.65	87.37	0.200	9.04
	R40	42.98	0.216	15.90	77.47	0.196	19.29
	R60	37.77	0.209	24.08	64.33	0.189	29.40
	R80	34.34	0.197	32.31	53.58	0.179	39.55

Table 30: **Pruning patterns on Qwen3-4B-Thinking-2507.** Task metrics, warm-cache completion time (t_{TC}), and pruning ratio (ρ) are reported.

Category	Method	LongBench-v2			SQuAD v2			
		$Q_{abs}\uparrow$	$t_{TC}\downarrow$	ρ	EM ↑	F1 ↑	$t_{TC}\downarrow$	ρ
	Baseline	14.00	255.749	–	37.00	54.29	31.572	–
Structural Subword	S1	6.00	202.696	53.01	37.00	46.41	55.971	24.90
	S2	10.00	197.447	45.95	42.00	54.19	39.075	18.93
Punctuation	P1	10.00	236.791	24.73	44.00	62.77	34.214	10.27
Lexical & POS	L1	6.00	212.697	36.75	28.00	38.11	59.544	43.59
	L2	12.00	213.376	31.89	30.00	40.55	50.396	33.78
Frequency	F1	6.00	216.606	48.56	26.00	27.50	35.681	68.42
	F2	12.00	209.648	56.79	39.00	39.47	29.118	80.20
Syntactic	D1	18.00	145.369	86.14	33.00	34.89	36.724	75.80
	D2	18.00	134.541	89.11	31.00	32.40	28.175	86.40
Morphology	M1	12.00	222.645	52.90	35.00	43.93	58.116	24.53
	M2	6.00	230.647	41.81	36.00	52.68	45.647	13.44
Random Reference	R20	10.00	245.808	25.54	30.00	39.45	61.705	19.54
	R40	8.00	194.192	44.13	21.00	26.75	65.277	39.63
	R60	12.00	222.474	62.71	34.00	38.22	61.177	59.59
	R80	8.00	147.867	81.36	41.00	42.72	34.138	79.73

Category	Method	LooGLE			QASPER			
		ROUGE-L ↑	$t_{TC}\downarrow$	ρ	EM ↑	F1 ↑	$t_{TC}\downarrow$	ρ
	Baseline	22.14	108.539	–	10.00	47.82	49.037	–
Structural Subword	S1	12.04	141.139	32.91	15.00	45.76	56.334	25.94
	S2	12.28	141.202	24.90	12.00	48.02	54.146	19.21
Punctuation	P1	22.82	106.791	15.61	12.00	46.27	44.250	12.28
Lexical & POS	L1	17.73	110.554	46.65	4.00	23.26	55.359	42.18
	L2	18.02	123.434	33.23	6.00	35.58	48.468	32.38
Frequency	F1	7.00	90.107	67.17	6.00	11.63	44.523	66.24
	F2	4.54	88.020	79.23	9.00	9.86	33.847	76.55
Syntactic	D1	6.66	103.027	77.81	5.00	12.81	54.781	77.22
	D2	5.82	102.507	79.29	12.00	17.34	53.645	84.68
Morphology	M1	10.90	138.525	32.76	13.00	40.76	57.526	25.52
	M2	13.54	121.367	21.90	9.00	44.15	46.262	17.51
Random Reference	R20	18.32	121.078	22.81	10.00	39.08	50.254	20.34
	R40	11.78	158.019	42.18	9.00	29.70	59.853	40.35
	R60	6.20	121.457	61.42	11.00	20.66	62.880	60.17
	R80	3.12	90.501	80.72	12.00	15.75	41.220	80.05

Category	Method	MMLU-Pro			ARC-Challenge		
		$Q_{abs}\uparrow$	$t_{TC}\downarrow$	ρ	$Q_{abs}\uparrow$	$t_{TC}\downarrow$	ρ
	Baseline	19.00	105.986	–	36.00	45.397	–
Structural Subword	S1	19.00	122.569	10.43	39.00	48.540	6.38
	S2	17.00	114.617	9.46	39.00	46.255	4.95
Punctuation	P1	20.00	110.449	4.08	37.00	46.510	3.95
Lexical & POS	L1	13.00	127.101	14.94	49.00	63.472	25.45
	L2	20.00	120.438	11.09	48.00	54.436	17.84
Frequency	F1	12.00	119.854	24.71	13.00	53.780	40.30
	F2	13.00	112.590	28.47	10.00	36.639	46.98
Syntactic	D1	20.00	116.874	27.91	22.00	60.965	36.12
	D2	17.00	102.429	30.17	21.00	62.349	37.75
Morphology	M1	17.00	123.008	10.37	39.00	49.169	6.26
	M2	18.00	117.814	6.97	44.00	49.945	4.96
Random Reference	R20	14.00	119.606	6.58	48.00	54.471	8.89
	R40	14.00	136.895	14.26	30.00	54.630	18.68
	R60	15.00	126.783	21.48	27.00	57.348	28.28
	R80	12.00	116.601	28.80	19.00	56.980	38.32

D LONG-CONTEXT ANALYSIS AND THE COVERAGE PARADOX

This section provides the complete evidence for the coverage paradox discussed in Section 4.4.

Truncation incidence. Table 31 reports the fraction of inputs that exceed the configured context length before and after pruning. On LongBench-v2, 36.2% (Llama) to 94.4% (Phi-4) of baseline inputs are truncated. Conservative patterns reduce this fraction only moderately, whereas the syntactic patterns reduce it to 0–47% (D2: 2.8% for the Llama models, 17.7% for Qwen3.5-9B, and 39.8% for Phi-4). On LooGLE, the syntactic and frequency rules eliminate truncation almost completely for all models, while QASPER, SQuAD v2, MMLU-Pro, and ARC-Challenge are essentially unaffected by truncation (Tab. 8). Because truncated baseline inputs already fill the context window, pruning does not reduce the peak memory on LongBench-v2 (App. E); instead, the same token budget is filled with a larger portion of the source document.

Table 31: **Truncation incidence** (% of inputs that still exceed the configured context length after pruning) on LongBench-v2 and LooGLE. Pruning before tokenization shifts inputs below the context boundary, so that more of the source document reaches the model. †Evaluated on subsets (100 samples per benchmark; 50 on LongBench-v2).

	LongBench-v2							LooGLE						
Model	Base	P1	L2	F2	D1	D2	R80	Base	P1	L2	F2	D1	D2	R80
Llama-3.2-3B	36.2	28.6	23.5	12.3	3.8	2.8	7.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Llama-3.1-8B	36.2	28.6	23.5	12.3	3.8	2.8	7.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mistral-7B	80.3	74.0	71.2	47.3	26.4	22.7	33.6	58.1	37.4	16.3	0.0	0.0	0.0	0.0
Qwen2.5-7B	76.9	72.2	68.4	45.7	24.2	19.1	30.2	50.9	20.7	6.9	0.0	0.0	0.0	0.0
Qwen2.5-14B	76.9	72.2	68.4	45.7	24.2	19.1	30.2	50.9	20.7	6.9	0.0	0.0	0.0	0.0
Qwen3.5-4B	77.9	72.4	68.4	45.7	24.2	17.7	30.2	52.9	18.8	6.9	0.0	0.0	0.0	0.0
Qwen3.5-9B	77.9	72.4	68.4	45.7	24.2	17.7	30.2	52.9	18.8	6.9	0.0	0.0	0.0	0.0
Phi-4	94.4	86.5	82.1	60.0	46.7	39.8	55.7	83.5	75.3	65.8	0.9	0.0	0.4	0.0
Qwen3-4B-Think.†	22.0	18.0	16.0	10.0	0.0	0.0	4.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Length stratification. LongBench-v2 assigns each item to one of three length categories (short: n=180, medium: n=215, long: n=108), which we use as a proxy for context pressure. Table 33 and Figures 8–14 report relative accuracy per category for all non-reasoning models. For D2, relative accuracy increases from the short to the long category for every model; it reaches or exceeds the baseline on long documents for six of eight models, with Llama-3.1-8B (0.93) and Qwen3.5-4B (0.97) as exceptions. The reversal is most pronounced for Phi-4 and Mistral-7B, whose baselines truncate the largest fraction of LongBench-v2 inputs. The conservative rule P1 is at or above the baseline on long documents for seven of eight models. Given the category sizes, differences of about 0.1 in relative accuracy correspond to only a few items; we therefore base our conclusions on the consistency of the pattern across models and on the paired tests (App. C.5) rather than on individual cells.

Context budget as a natural experiment. All models are evaluated on the same LongBench-v2 items, but with different effective context budgets. Table 32 relates each model's baseline truncation incidence to the change in D2's relative accuracy from short to long documents. The rank correlation is ρ_s=0.93 for D2 and 0.84 for D1, but only 0.17 for R80 and 0.24 for F1. Because three pairs of models share identical truncation rates, only five distinct budgets are represented, and the correlation should be read as descriptive. Absolute counts show that the largest relative gains correspond to few items (e.g., Phi-4: 15 $\rightarrow$ 30 of 108).

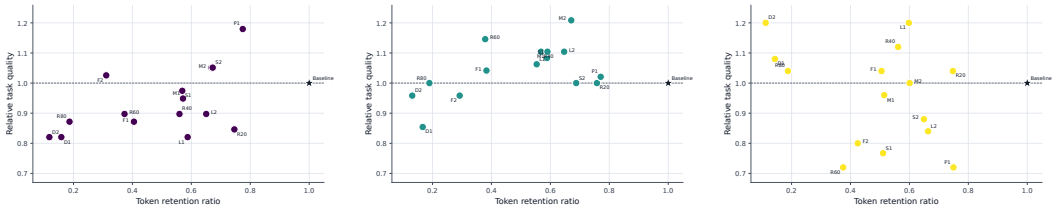
Why the paradox does not appear on LooGLE. Although truncation affects 50.9–83.5% of LooGLE baseline inputs for all models except the two Llama models, all interventions reduce ROUGE-L for Qwen3.5-9B, and only small, conservative rules yield significant gains for Qwen2.5-14B. LooGLE LongDepQA requires free-form answers that are scored by lexical overlap with a reference, so the loss of local detail — which directly reduces the overlap of generated answers — outweighs the recovered coverage. LongBench-v2, in contrast, is a multiple-choice task for which telegraphic but complete evidence can suffice. The coverage paradox therefore depends on context pressure *and* on the granularity of evidence and output required by the task.

Table 32: **Coverage paradox across context budgets** (D2, LongBench-v2). Baseline truncation incidence, relative accuracy on short and long documents, and correct long-document items (baseline → D2, out of 108).

Model	Truncated (%)	Short	Long	Long − Short	Correct (long)
Llama-3.2-3B	36.2	0.82	1.20	+0.38	25 → 30
Llama-3.1-8B	36.2	0.74	0.93	+0.19	28 → 26
Qwen2.5-7B	76.9	0.78	1.07	+0.29	29 → 31
Qwen2.5-14B	76.9	0.69	1.09	+0.40	32 → 35
Qwen3.5-4B	77.9	0.54	0.97	+0.44	37 → 36
Qwen3.5-9B	77.9	0.69	1.10	+0.41	30 → 33
Mistral-7B	80.3	0.90	1.38	+0.48	21 → 29
Phi-4	94.4	0.84	2.00	+1.16	15 → 30
Spearman ρ_s (truncation vs. long − short): D2 0.93, D1 0.84, P1 0.35, F1 0.24, R80 0.17					

Table 33: **Relative accuracy by LongBench-v2 length category for all non-reasoning models** (short / medium / long). The reasoning model is omitted because its 50-sample LongBench-v2 subset leaves fewer than ten items per category.

Model	P1	F2	D1	D2	R80
Llama-3.2-3B	1.18 / 1.02 / 0.72	1.03 / 0.96 / 0.80	0.82 / 0.85 / 1.08	0.82 / 0.96 / 1.20	0.87 / 1.00 / 1.04
Llama-3.1-8B	0.90 / 0.73 / 1.07	0.78 / 0.73 / 1.04	0.78 / 0.78 / 0.79	0.74 / 0.88 / 0.93	0.74 / 0.84 / 1.00
Mistral-7B	1.06 / 1.09 / 1.05	0.92 / 0.98 / 1.19	0.82 / 1.05 / 1.43	0.90 / 0.98 / 1.38	0.90 / 0.75 / 1.00
Qwen2.5-7B	1.05 / 1.08 / 1.00	0.84 / 0.98 / 1.03	0.84 / 0.97 / 1.00	0.78 / 0.92 / 1.07	0.78 / 0.98 / 1.00
Qwen2.5-14B	1.01 / 1.06 / 1.25	0.69 / 0.90 / 1.06	0.59 / 0.99 / 0.97	0.69 / 0.91 / 1.09	0.65 / 0.88 / 1.00
Qwen3.5-4B	0.89 / 1.19 / 1.05	0.72 / 0.97 / 0.81	0.62 / 1.02 / 0.92	0.54 / 1.05 / 0.97	0.65 / 1.05 / 0.73
Qwen3.5-9B	0.99 / 1.13 / 1.10	0.48 / 0.83 / 0.93	0.56 / 0.99 / 0.87	0.69 / 1.04 / 1.10	0.54 / 0.75 / 1.03
Phi-4	1.14 / 1.09 / 1.73	1.20 / 0.83 / 1.13	1.06 / 1.20 / 2.07	0.84 / 1.20 / 2.00	0.82 / 0.87 / 1.87
Median	1.03 / 1.09 / 1.06	0.81 / 0.93 / 1.04	0.80 / 0.99 / 0.98	0.76 / 0.97 / 1.10	0.76 / 0.88 / 1.00



(a) Short. (b) Medium. (c) Long.

Figure 8: **LongBench-v2 results by length category for Llama-3.2-3B-Instruct.** Relative task accuracy over token retention ratio ($1 - \rho$) for the short, medium, and long documents.



(a) Short. (b) Medium. (c) Long.

Figure 9: **LongBench-v2 results by length category for Llama-3.1-8B-Instruct.** Relative task accuracy over token retention ratio ($1 - \rho$) for the short, medium, and long documents.



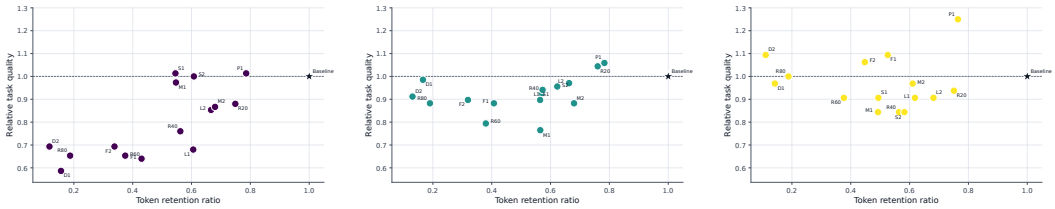
(a) Short. (b) Medium. (c) Long.

Figure 10: **LongBench-v2 results by length category for Mistral-7B-Instruct-v0.3.** Relative task accuracy over token retention ratio $(1 - \rho)$ for the short, medium, and long documents.



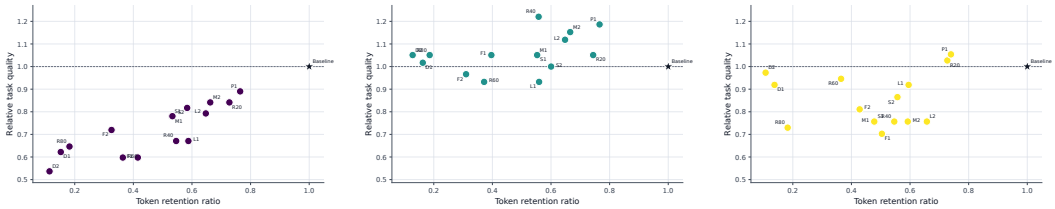
(a) Short. (b) Medium. (c) Long.

Figure 11: **LongBench-v2 results by length category for Qwen2.5-7B-Instruct.** Relative task accuracy over token retention ratio $(1 - \rho)$ for the short, medium, and long documents.



(a) Short. (b) Medium. (c) Long.

Figure 12: **LongBench-v2 results by length category for Qwen2.5-14B-Instruct.** Relative task accuracy over token retention ratio $(1 - \rho)$ for the short, medium, and long documents.



(a) Short. (b) Medium. (c) Long.

Figure 13: **LongBench-v2 results by length category for Qwen3.5-4B.** Relative task accuracy over token retention ratio $(1 - \rho)$ for the short, medium, and long documents.



(a) Short. (b) Medium. (c) Long.

Figure 14: **LongBench-v2 results by length category for Phi-4.** Relative task accuracy over token retention ratio ($1 - \rho$) for the short, medium, and long documents.

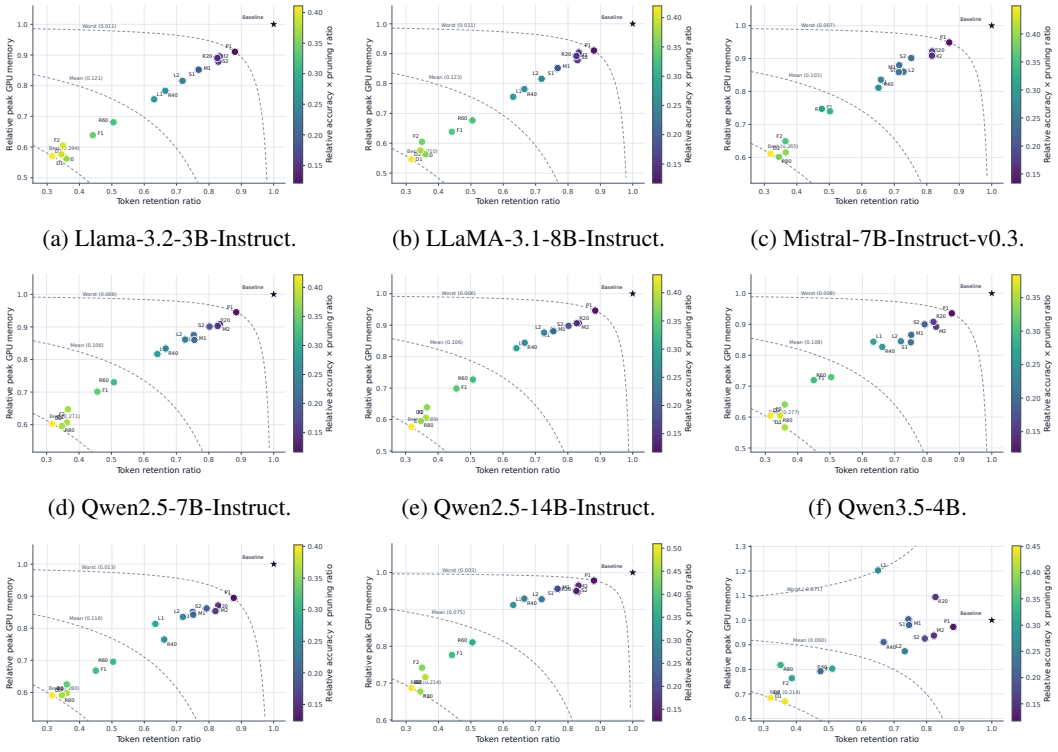
E PEAK GPU MEMORY

We measure the peak GPU memory allocated during inference and subtract the static footprint of the loaded model, so that the reported values isolate the memory required for processing the input (activations and KV cache). Table 34 reports the relative reduction for every model and intervention, and Figure 15 relates it to the retention ratio.

Memory scales with reduction and input length. Averaged over the six benchmarks, the syntactic patterns D1 and D2 and the random reference R80 yield the largest reductions (37–39% on average over all cells; Tab. 9), followed by F2 (34%), whereas conservative patterns save 6–14%. The reductions are much larger on long inputs that fit into the window: on QASPER, D2 reduces the input-dependent peak memory by 73–90% across models (89.7% for Qwen3.5-9B). On LongBench-v2, in contrast, memory remains nearly unchanged for most models because truncated inputs fill the context window both before and after pruning; here, pruning trades memory savings for coverage (App. D). On the short inputs of MMLU-Pro and ARC-Challenge, absolute memory differences are small, so relative values are less informative. For the reasoning model, memory is dominated by the reasoning trace, and several interventions increase peak memory (negative values in Tab. 34).

Table 34: **Relative reduction of peak GPU memory during inference** (1 − pruned/baseline peak allocation above the static model footprint), macro-averaged over the six benchmarks. †Evaluated on subsets (100 samples per benchmark; 50 on LongBench-v2).

Model	P1	S1	S2	L1	L2	F1	F2	D1	D2	M1	M2	R20	R40	R60	R80
Llama-3.2-3B	0.089	0.147	0.105	0.244	0.184	0.361	0.395	0.438	0.429	0.148	0.122	0.110	0.216	0.319	0.424
Llama-3.1-8B	0.089	0.150	0.096	0.245	0.184	0.361	0.395	0.437	0.454	0.148	0.122	0.108	0.219	0.323	0.424
Mistral-7B	0.051	0.120	0.098	0.189	0.140	0.253	0.351	0.385	0.389	0.141	0.078	0.091	0.164	0.260	0.399
Qwen2.5-7B	0.055	0.125	0.099	0.183	0.139	0.299	0.353	0.393	0.397	0.140	0.091	0.097	0.166	0.270	0.404
Qwen2.5-14B	0.054	0.116	0.102	0.173	0.124	0.301	0.361	0.394	0.423	0.119	0.093	0.094	0.156	0.273	0.405
Qwen3.5-4B	0.065	0.158	0.100	0.156	0.155	0.281	0.359	0.433	0.395	0.134	0.108	0.092	0.173	0.271	0.395
Qwen3.5-9B	0.105	0.149	0.138	0.186	0.165	0.332	0.375	0.402	0.409	0.158	0.129	0.146	0.235	0.304	0.409
Phi-4	0.022	0.045	0.035	0.088	0.073	0.224	0.258	0.284	0.313	0.044	0.053	0.050	0.071	0.189	0.323
Qwen3-4B-Think.†	0.027	-0.004	0.075	-0.203	0.126	0.208	0.236	0.330	0.316	0.020	0.063	-0.094	0.089	0.197	0.182
Mean	0.062	0.112	0.094	0.140	0.143	0.291	0.343	0.389	0.392	0.117	0.095	0.077	0.166	0.268	0.374



(a) Llama-3.2-3B-Instruct. (b) LLaMA-3.1-8B-Instruct. (c) Mistral-7B-Instruct-v0.3.

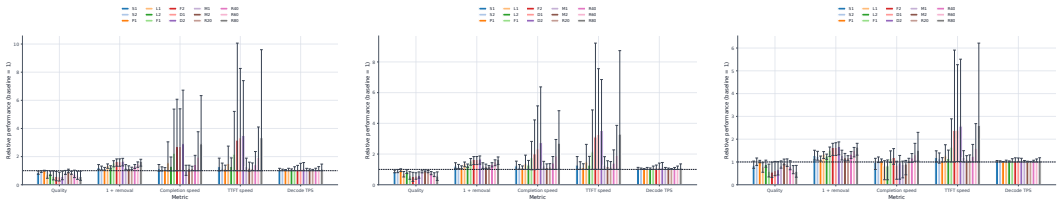
(d) Qwen2.5-7B-Instruct. (e) Qwen2.5-14B-Instruct. (f) Qwen3.5-4B.

(g) Qwen3.5-9B. (h) Phi-4. (i) Qwen3-4B-Thinking-2507.

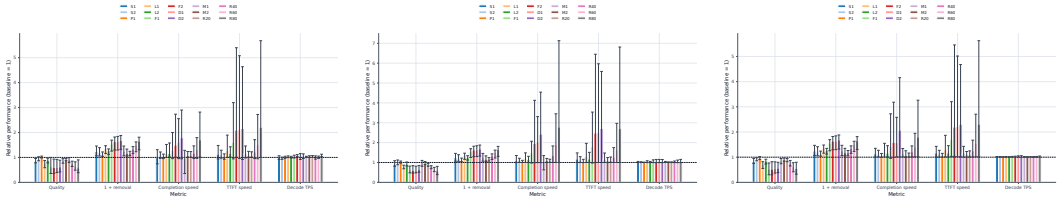
Figure 15: **Peak GPU memory under static pruning.** Relative peak memory (above the static model footprint) over token retention ratio ($1 - \rho$); color encodes saved memory multiplied by the pruning ratio (higher is better). All values are normalized to the unpruned baseline.

F Relative Quality and Efficiency Metrics per Model

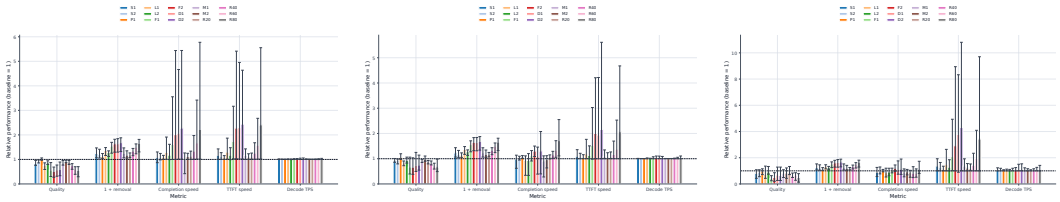
Figure 16 summarizes, for every model, the relative quality, the token removal, the end-to-end and time-to-first-token speed-ups, and the decoding throughput of each intervention, averaged over the six benchmarks, with whiskers spanning the range over benchmarks. Decoding throughput remains close to the baseline for all models and interventions, since the model and its weights are unchanged. The efficiency gains therefore arise from the shorter prefill, and the wide ranges of completion and TTFT speed reflect the dependence of these gains on input length: they are large on long-document benchmarks and negligible on the short inputs of MMLU-Pro and ARC-Challenge.



(a) Llama-3.2-3B-Instruct. (b) LLaMA-3.1-8B-Instruct. (c) Mistral-7B-Instruct-v0.3.



(d) Qwen2.5-7B-Instruct. (e) Qwen2.5-14B-Instruct. (f) Qwen3.5-4B.



(g) Qwen3.5-9B. (h) Phi-4. (i) Qwen3-4B-Thinking-2507.

Figure 16: **Aggregate relative metrics across all evaluated benchmarks.** Bar heights represent the mean relative to the unpruned baseline (1.0, dashed line), and whiskers show the range (min–max) across individual benchmarks. Metrics report relative task quality, token removal (1 + removal), warm-cache completion speed, time-to-first-token (TTFT) speed, and decode throughput (TPS).

G MULTILINGUAL TRANSFER

The main experiments characterize the interventions on English benchmarks. To test whether the observed sensitivities transfer beyond English, we conduct a controlled study on XQuAD (Artetxe et al., 2020) with Llama-3.2-3B-Instruct, Qwen2.5-7B-Instruct, and the discovery model Qwen3.5-9B. The experiment characterizes *where* the rule families transfer and where they fail; it is not intended to establish universal multilingual invariance.

Setup. We evaluate the same 1,190 XQuAD questions in English, German, and Turkish. German provides a fusional comparison, while Turkish is agglutinative and encodes substantially more grammatical information within individual word forms. We use language-specific frequency resources and parsers (Stanza for Turkish dependency parsing). Because baseline difficulty differs across languages, each intervention is evaluated relative to its same-language unpruned baseline. We report token-level F1 as the primary metric. Exact match is zero for nearly all XQuAD configurations because model-specific end-of-turn tokens (e.g., `<|im_end|>`, `<|eot_id|>`) were not stripped from decoded answers before scoring. This affects exact match but only marginally lowers F1, and it affects all interventions and the baseline equally, so relative F1 remains comparable. Tables 35–37 report absolute results, and Figures 17–22 visualize them. All XQuAD contexts fit into the context window, so this study probes the short-context regime.

Conservative surface transformations transfer most reliably. P1 remains close to the unpruned baseline in all three languages, retaining 99.2%, 97.6%, and 97.7% of the English, German, and Turkish baseline F1 of Qwen3.5-9B while removing 8–10% of prompt tokens (91.0–99.2% across all three models). Stemming (M2) retains 78–82% for Qwen3.5-9B in all three languages. Conservative surface-level redundancy is thus comparatively stable across the evaluated languages.

Function-word sensitivity depends on linguistic structure. L2 removes 33.2% of English prompt tokens for Qwen3.5-9B but only 12.6% in Turkish, where it retains 91.7% of the baseline F1 (English: 81.0%). This is consistent with grammatical information being encoded within Turkish word forms rather than in separate function words. Consequently, neither the achievable reduction nor the functional effect of an English-derived function-word rule transfers unchanged.

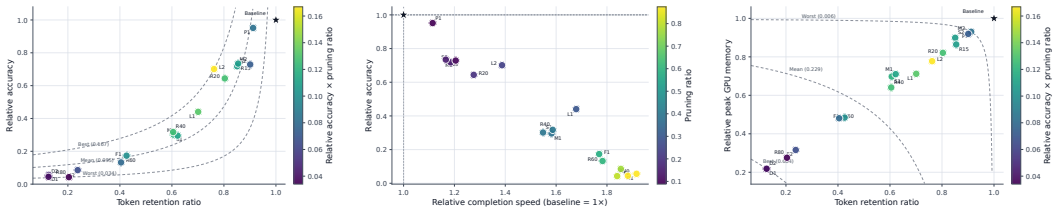
Morphological and subword reduction transfers poorly. M1 retains 58.5% of the English F1 of Qwen3.5-9B but only 26.4% and 10.4% in German and Turkish, while its token removal increases from 25.4% (English) to 38.7% (German) and 49.7% (Turkish); S1 behaves almost identically. In morphologically richer languages, truncating a word form to a single subword removes case, number, person, tense, or lexical identity rather than redundant material. The same pattern holds for all three models.

Parser-dependent rules can fail entirely. For German, D1 removes the entire context (ρ=100%) and D2 removes 95.5–95.8% for all three models, which reduces F1 to zero. The rules retain words by dependency label, and the retained labels presumably do not occur in the annotation scheme of the German pipeline. This is a failure of rule transfer rather than evidence about German syntax, and it illustrates that parser-based rules require language-specific validation. The aggressive rules F2, D1, and D2 also collapse in English and Turkish on this short-context task (relative F1 of at most 0.13), consistent with the behavior on the short-context benchmarks in the main study.

Structured reduction versus random deletion. To test whether behavior is explained by sequence length alone, we compare S2 with random deletion at similar rates (R15, R20). In English, S2 exceeds the random reference for all three models (e.g., Qwen3.5-9B: 0.76 at ρ=19.9% versus 0.62 for R20 at ρ=19.6%). In German and Turkish, S2 is on par with or below random deletion for Llama-3.2-3B and Qwen2.5-7B and only marginally above it for Qwen3.5-9B. The advantage of structured subword reduction is therefore itself language-dependent.

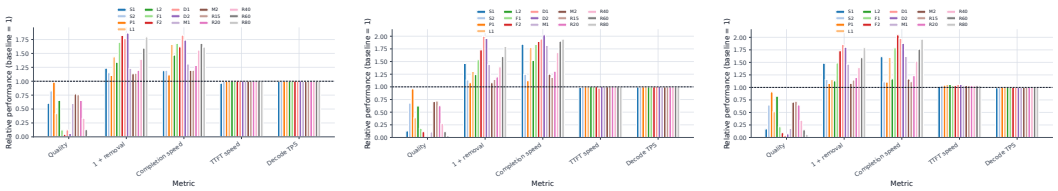
Takeaway. These experiments provide evidence for *family-specific multilingual transfer*, not universal linguistic stability. Conservative surface rules transfer across the evaluated languages, whereas morphology-, parser-, and frequency-dependent rules require language-specific resources, rule selection, and validation. Rankings of interventions remain correlated across models within each

language (Spearman ρ_s between 0.73 and 0.96). We therefore restrict the main generalization claims to English and treat multilingual transfer as an explicit boundary of the characterization.



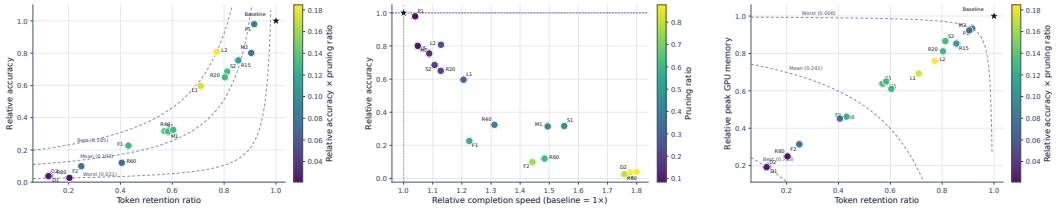
(a) Relative task accuracy. (b) Relative completion speed. (c) Relative peak memory.

Figure 17: **Multilingual transfer of Llama-3.2-3B-Instruct on XQuAD.** Relative F1, completion speed, and peak memory aggregated over English, German, and Turkish.



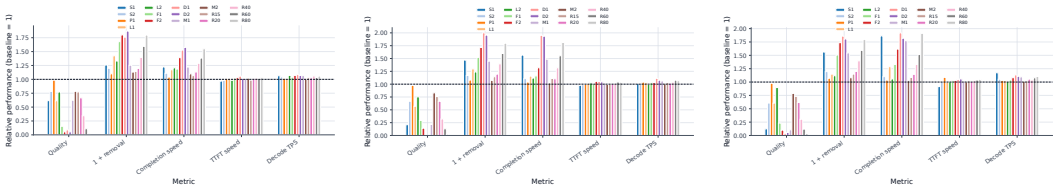
(a) English. (b) German. (c) Turkish.

Figure 18: **Multilingual transfer of Llama-3.2-3B-Instruct on XQuAD, per language.** Bars show relative quality, reduction, and efficiency metrics for each intervention.



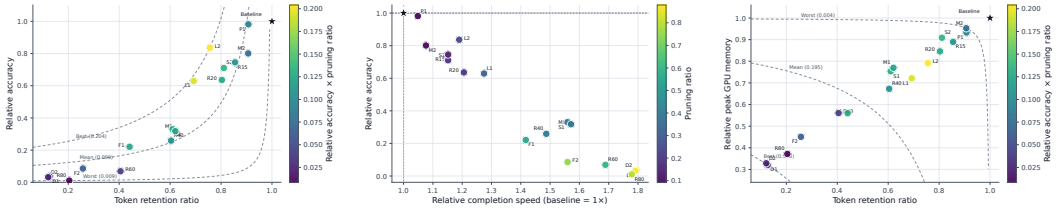
(a) Relative task accuracy. (b) Relative completion speed. (c) Relative peak memory.

Figure 19: **Multilingual transfer of Qwen2.5-7B-Instruct on XQuAD.** Relative F1, completion speed, and peak memory aggregated over English, German, and Turkish.



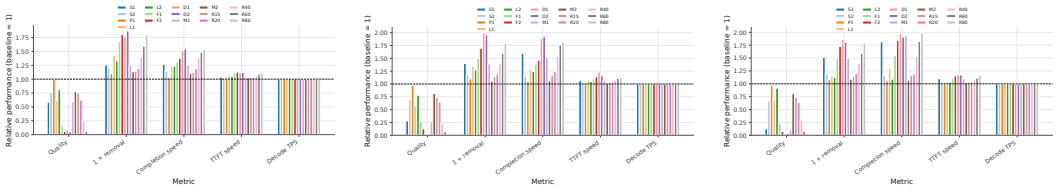
(a) English. (b) German. (c) Turkish.

Figure 20: **Multilingual transfer of Qwen2.5-7B-Instruct on XQuAD, per language.** Bars show relative quality, reduction, and efficiency metrics for each intervention.



(a) Relative task accuracy. (b) Relative completion speed. (c) Relative peak memory.

Figure 21: **Multilingual transfer of Qwen3.5-9B on XQuAD.** Relative F1, completion speed, and peak memory aggregated over English, German, and Turkish.



(a) English. (b) German. (c) Turkish.

Figure 22: **Multilingual transfer of Qwen3.5-9B on XQuAD, per language.** Bars show relative quality, reduction, and efficiency metrics for each intervention.

Table 35: **Multilingual XQuAD pruning on Llama-3.2-3B-Instruct.** Task metrics, warm-cache completion time (t_{TC}), and pruning ratio (ρ) are reported.

Category	Method	XQuAD (EN)				XQuAD (DE)			
		EM ↑	F1 ↑	t_{TC} ↓	ρ	EM ↑	F1 ↑	t_{TC} ↓	ρ
	Baseline	0.00	62.01	0.165	–	0.00	52.01	0.198	–
Structural Subword	S1	0.00	37.45	0.139	23.48	0.00	6.71	0.108	46.18
	S2	0.00	51.24	0.138	15.32	0.00	35.32	0.159	13.65
Punctuation	P1	0.00	60.96	0.148	10.29	0.00	49.96	0.177	8.14
Lexical & POS	L1	0.00	25.88	0.099	43.82	0.00	20.26	0.112	30.33
	L2	0.00	40.67	0.112	34.12	0.00	32.38	0.130	24.18
Frequency	F1	0.00	7.72	0.098	69.96	0.00	9.43	0.108	53.48
	F2	0.00	2.82	0.102	82.34	0.00	5.97	0.105	73.13
Syntactic	D1	0.00	7.80	0.090	76.65	0.00	0.00	0.102	100.00
	D2	0.00	3.73	0.095	86.62	0.00	0.06	0.098	95.54
Morphology	M1	0.00	37.15	0.126	23.00	0.00	5.66	0.109	44.03
	M2	0.00	47.82	0.138	13.43	0.00	36.88	0.158	8.15
Random Reference	R15	0.00	46.98	0.138	14.37	0.00	37.52	0.167	14.49
	R20	0.00	40.60	0.129	19.63	0.00	32.64	0.151	19.62
	R40	0.00	20.73	0.106	39.45	0.00	14.29	0.118	39.57
	R60	0.00	8.02	0.098	59.52	0.00	5.96	0.104	59.79
	R80	0.00	1.59	0.102	79.72	0.00	2.04	0.102	79.78

Category	Method	XQuAD (TR)			
		EM ↑	F1 ↑	t_{TC} ↓	ρ
	Baseline	0.08	48.56	0.211	–
Structural Subword	S1	0.00	8.31	0.131	48.32
	S2	0.00	31.50	0.190	15.79
Punctuation	P1	0.00	44.21	0.191	7.91
Lexical & POS	L1	0.00	24.99	0.132	15.56
	L2	0.00	40.06	0.181	13.10
Frequency	F1	0.00	10.43	0.118	48.89
	F2	0.00	4.65	0.103	73.23
Syntactic	D1	0.00	2.11	0.107	85.53
	D2	0.00	3.50	0.112	80.19
Morphology	M1	0.00	8.69	0.131	46.40
	M2	0.00	34.30	0.181	8.13
Random Reference	R15	0.00	35.16	0.190	14.73
	R20	0.08	31.53	0.171	19.73
	R40	0.00	16.66	0.139	39.55
	R60	0.00	7.42	0.120	59.36
	R80	0.00	3.07	0.108	79.35

Table 36: **Multilingual XQuAD pruning on Qwen2.5-7B-Instruct.** Task metrics, warm-cache completion time (t_{TC}), and pruning ratio (ρ) are reported.

Category	Method	XQuAD (EN)				XQuAD (DE)			
		EM ↑	F1 ↑	t_{TC} ↓	ρ	EM ↑	F1 ↑	t_{TC} ↓	ρ
	Baseline	0.00	65.69	0.533	–	0.08	59.48	0.653	–
Structural Subword	S1	0.00	40.47	0.436	25.84	0.00	12.49	0.417	46.97
	S2	0.00	51.37	0.480	19.94	0.00	39.66	0.589	16.86
Punctuation	P1	0.00	64.73	0.512	9.97	0.00	58.06	0.627	8.03
Lexical & POS	L1	0.00	40.34	0.455	42.43	0.08	33.90	0.564	29.80
	L2	0.00	50.74	0.441	33.02	0.00	44.78	0.583	23.79
Frequency	F1	0.00	10.17	0.450	67.90	0.00	17.36	0.562	52.23
	F2	0.00	3.54	0.383	80.06	0.00	8.46	0.495	71.55
Syntactic	D1	0.00	5.87	0.349	76.06	0.00	0.00	0.335	100.00
	D2	0.00	3.53	0.338	86.99	0.00	0.03	0.337	95.64
Morphology	M1	0.00	40.95	0.436	25.35	0.00	12.74	0.439	44.83
	M2	0.00	51.51	0.485	13.16	0.00	49.43	0.639	7.40
Random Reference	R15	0.00	50.82	0.500	14.28	0.00	45.11	0.588	14.46
	R20	0.00	43.82	0.470	19.64	0.00	39.67	0.591	19.59
	R40	0.00	22.65	0.414	39.49	0.00	19.19	0.495	39.61
	R60	0.00	7.26	0.385	59.43	0.00	7.74	0.420	59.77
	R80	0.00	1.23	0.344	79.77	0.00	1.44	0.360	79.77

Category	Method	XQuAD (TR)			
		EM ↑	F1 ↑	t_{TC} ↓	ρ
	Baseline	0.00	55.52	0.746	–
Structural Subword	S1	0.00	6.87	0.400	56.22
	S2	0.00	33.73	0.677	19.57
Punctuation	P1	0.00	54.23	0.719	7.17
Lexical & POS	L1	0.00	33.64	0.579	14.85
	L2	0.00	49.81	0.705	11.81
Frequency	F1	0.00	12.89	0.561	50.64
	F2	0.00	5.57	0.461	73.72
Syntactic	D1	0.00	1.78	0.388	85.94
	D2	0.00	3.13	0.410	80.68
Morphology	M1	0.00	5.85	0.420	54.56
	M2	0.00	43.74	0.727	8.11
Random Reference	R15	0.00	40.68	0.686	14.87
	R20	0.00	34.24	0.652	19.91
	R40	0.00	16.88	0.562	39.78
	R60	0.00	6.62	0.492	59.30
	R80	0.00	2.00	0.390	79.48

Table 37: **Multilingual XQuAD pruning on Qwen3.5-9B.** Task metrics, warm-cache completion time (t_{TC}), and pruning ratio (ρ) are reported.

Category	Method	XQuAD (EN)				XQuAD (DE)			
		EM ↑	F1 ↑	t_{TC} ↓	ρ	EM ↑	F1 ↑	t_{TC} ↓	ρ
	Baseline	0.00	66.13	0.266	–	0.00	62.63	0.321	–
Structural Subword	S1	0.00	38.73	0.211	25.76	0.00	17.59	0.201	40.06
	S2	0.00	50.56	0.234	19.94	0.00	43.65	0.278	17.48
Punctuation	P1	0.00	65.59	0.258	10.01	0.00	61.15	0.305	9.76
Lexical & POS	L1	0.00	40.91	0.215	42.61	0.00	35.80	0.251	34.45
	L2	0.00	53.57	0.215	33.16	0.00	49.00	0.258	27.42
Frequency	F1	0.00	10.46	0.203	68.16	0.00	17.29	0.231	50.11
	F2	0.00	3.73	0.194	80.26	0.00	7.91	0.219	69.78
Syntactic	D1	0.00	6.34	0.175	76.09	0.00	0.00	0.170	100.00
	D2	0.00	3.50	0.173	87.01	0.00	0.03	0.167	95.77
Morphology	M1	0.00	38.70	0.212	25.36	0.00	16.55	0.211	38.72
	M2	0.00	51.34	0.241	13.34	0.00	51.26	0.302	5.66
Random Reference	R15	0.00	49.96	0.236	14.27	0.00	46.17	0.274	14.46
	R20	0.00	41.26	0.225	19.63	0.00	40.62	0.258	19.53
	R40	0.00	16.39	0.191	39.50	0.00	14.99	0.208	39.68
	R60	0.00	3.83	0.179	59.46	0.00	4.25	0.182	59.74
	R80	0.00	0.60	0.174	79.74	0.00	0.55	0.177	79.77

Category	Method	XQuAD (TR)			
		EM ↑	F1 ↑	t_{TC} ↓	ρ
	Baseline	0.00	58.96	0.356	–
Structural Subword	S1	0.00	7.34	0.196	51.24
	S2	0.00	39.48	0.306	19.30
Punctuation	P1	0.00	57.58	0.335	8.02
Lexical & POS	L1	0.00	41.11	0.272	15.18
	L2	0.00	54.06	0.326	12.63
Frequency	F1	0.00	13.47	0.229	49.49
	F2	0.00	4.40	0.193	72.73
Syntactic	D1	0.00	0.72	0.180	85.85
	D2	0.00	2.44	0.186	80.86
Morphology	M1	0.00	6.10	0.184	49.74
	M2	0.00	47.65	0.335	8.97
Random Reference	R15	0.00	43.83	0.307	14.75
	R20	0.00	37.46	0.299	19.85
	R40	0.00	17.09	0.233	39.65
	R60	0.00	4.73	0.195	59.26
	R80	0.00	0.85	0.178	79.28

Table 38: **LLMLingua-2 results by model**, averaged over SQuAD v2, LongBench-v2, LooGLE, and QASPER. Q_{rel} is the score relative to the unpruned baseline of the same model, ρ the achieved token reduction, rel. t_{TC} the end-to-end speed-up (warm cache, excluding the compressor's own inference), and rel. M_{GPU} the relative reduction of peak GPU memory; higher is better in all four columns.

Model	Method	Q_{rel} ↑	ρ	rel. t_{TC} ↑	rel. M_{GPU} ↑
Llama-3.2-3B-Instruct	LLMLingua2-20pct	1.0152	0.1860	1.1580	0.2485
Llama-3.2-3B-Instruct	LLMLingua2-40pct	0.9777	0.3946	1.3612	0.3655
Llama-3.2-3B-Instruct	LLMLingua2-60pct	0.9323	0.5921	1.6471	0.5065
Llama-3.2-3B-Instruct	LLMLingua2-80pct	0.8522	0.7913	2.2683	0.5985
Qwen2.5-7B-Instruct	LLMLingua2-20pct	1.0153	0.1811	1.0210	0.0958
Qwen2.5-7B-Instruct	LLMLingua2-40pct	0.9566	0.3867	1.1620	0.2175
Qwen2.5-7B-Instruct	LLMLingua2-60pct	0.9531	0.5840	1.3301	0.3164
Qwen2.5-7B-Instruct	LLMLingua2-80pct	0.9098	0.7858	1.6689	0.5220
Qwen3.5-9B	LLMLingua2-20pct	1.0102	0.1782	1.0564	0.1572
Qwen3.5-9B	LLMLingua2-40pct	0.9944	0.3834	1.2332	0.2270
Qwen3.5-9B	LLMLingua2-60pct	0.9391	0.5785	1.4416	0.3743
Qwen3.5-9B	LLMLingua2-80pct	0.8147	0.7768	1.8448	0.5220

H Out-of-Scope Reference: LLMLingua-2 (Learned Compression)

This is not a baseline comparison. We include LLMLingua-2 (Pan et al., 2024) for completeness only, as an external reference point for what a *learned* prompt compressor achieves on the same benchmarks. It is *not* a like-for-like baseline for the research question of this paper, and the results below are *not* used as evidence for or against any of our findings. LLMLingua-2 trains a token classifier on compression data distilled from a strong LLM and decides, for each individual input, which tokens to keep. Our interventions, in contrast, are fixed rules that require no training and are held constant across inputs and models in order to measure whether predefined linguistic structures are load-bearing. A learned, input-adaptive compressor is therefore expected to retain substantially more utility at a given reduction level, and our rules were never designed to compete with it.

Setup. We run LLMLingua-2 with target reduction rates of 20, 40, 60, and 80% on the four context benchmarks (SQuAD v2, LongBench-v2, LooGLE, QASPER) for Llama-3.2-3B-Instruct, Qwen2.5-7B-Instruct, and Qwen3.5-9B. All other settings, including the target models, prompts, and metrics, are identical to the main study. The achieved token reduction closely follows the target (18–78%; Tab. 40). We use the checkpoint `microsoft/llmlingua-2-xlm-roberta-large-meetingbank` on the GPU. Compression is precomputed and cached, so the reported warm-cache completion times t_{TC} do not include the inference time of the compressor itself and therefore understate the end-to-end cost of LLMLingua-2 relative to our rules.

Observations. As expected, the learned compressor preserves considerably more utility than any fixed rule. For Qwen3.5-9B, LLMLingua-2 retains 93.9% of baseline quality at 57.9% reduction and 81.5% at 77.7% reduction (Tab. 38), compared with 58.8% for D2 at 68.4% reduction. Averaged over models, the gap at high reduction is largest on QASPER and LooGLE, where answers depend on local detail, and smallest on LongBench-v2 (Tab. 39; the corresponding values for the static rules are given per benchmark in Tabs. 22, 25, and 28). On LongBench-v2, LLMLingua-2 also scores above the unpruned baseline at moderate reduction (1.07× at 20% and 1.04× at 60% on average over the three models; Tab. 39), which is consistent with the coverage effect described in Section 4.4: whenever inputs are truncated, reduction admits additional evidence into the window, regardless of whether it is performed by a learned or a static method.

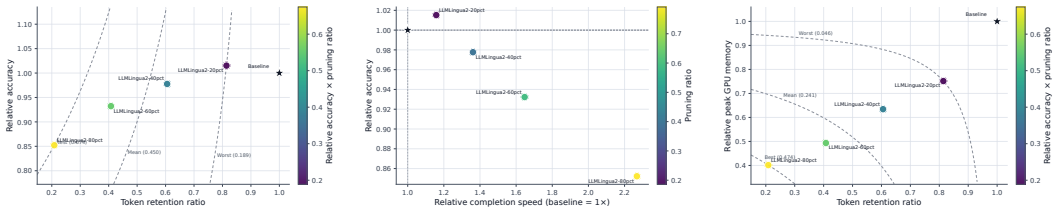
Interpretation. These results contextualize, but do not change, the contribution of this work. They confirm the expected ordering — adaptive learned compression retains more utility than fixed linguistic probes — while our rules require no model, no training data, and no per-input optimization, and they make the removed structure explicit and interpretable. A compressor that operationalizes the regularities identified in this study (e.g., prioritizing the syntactic core and punctuation-free surface forms under context pressure) would be the appropriate object for a direct comparison with LLMLingua-2; developing such a system is outside the scope of this characterization study.

Table 39: **LLMLingua-2 results by benchmark**, averaged over the three evaluated target models at each target reduction rate. Q_{rel} is the score relative to the unpruned baseline of the same model, ρ the achieved token reduction, rel. t_{TC} the end-to-end speed-up (warm cache, excluding the compressor's own inference), and rel. M_{GPU} the relative reduction of peak GPU memory. For all four columns, higher is better.

Dataset	Method	Q_{rel} ↑	ρ	rel. t_{TC} ↑	rel. M_{GPU} ↑
LongBench-v2	LLMLingua2-20pct	1.0734	0.1811	0.9984	-0.0005
LongBench-v2	LLMLingua2-40pct	1.0252	0.3870	1.2880	-0.0048
LongBench-v2	LLMLingua2-60pct	1.0443	0.5871	1.6074	0.0378
LongBench-v2	LLMLingua2-80pct	0.9750	0.7931	2.5376	0.0052
LooGLE	LLMLingua2-20pct	0.9925	0.1836	1.1385	0.1541
LooGLE	LLMLingua2-40pct	0.9877	0.3851	1.3309	0.2287
LooGLE	LLMLingua2-60pct	0.9708	0.5871	1.6145	0.3938
LooGLE	LLMLingua2-80pct	0.9030	0.7865	1.9799	0.6804
QASPER	LLMLingua2-20pct	1.0104	0.1757	1.1488	0.3337
QASPER	LLMLingua2-40pct	0.9808	0.3845	1.2888	0.5070
QASPER	LLMLingua2-60pct	0.9252	0.5835	1.4687	0.6677
QASPER	LLMLingua2-80pct	0.8374	0.7855	1.8950	0.8301
SQuAD v2	LLMLingua2-20pct	0.9779	0.1866	1.0283	0.1812
SQuAD v2	LLMLingua2-40pct	0.9112	0.3962	1.1008	0.3491
SQuAD v2	LLMLingua2-60pct	0.8256	0.5818	1.2012	0.4968
SQuAD v2	LLMLingua2-80pct	0.7203	0.7734	1.2969	0.6745

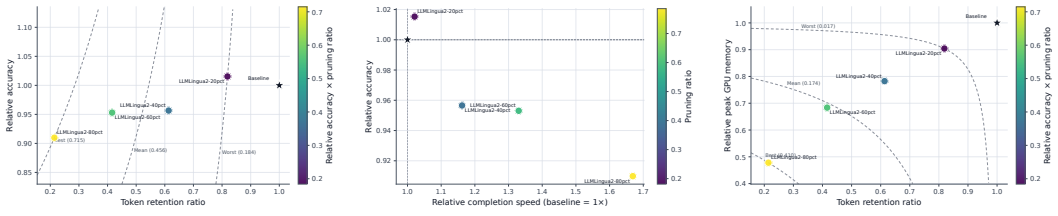
Table 40: **Calibration of LLMLingua-2: target versus achieved reduction.** ρ_{target} is the requested compression rate, ρ the token reduction actually measured with the target model's tokenizer, averaged over the three models and four benchmarks. Q_{rel} and rel. t_{TC} are the corresponding relative quality and end-to-end speed-up (higher is better).

Method	ρ_{target}	ρ	Q_{rel} ↑	rel. t_{TC} ↑
LLMLingua2-20pct	0.2000	0.1818	1.0136	1.0785
LLMLingua2-40pct	0.4000	0.3882	0.9762	1.2521
LLMLingua2-60pct	0.6000	0.5849	0.9415	1.4729
LLMLingua2-80pct	0.8000	0.7846	0.8589	1.9273



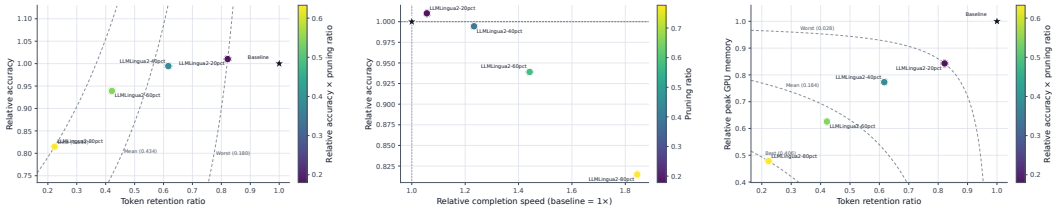
(a) Relative accuracy. (b) Relative completion speed. (c) Relative peak memory.

Figure 23: **LLMLingua-2 (learned, out-of-scope reference) on Llama-3.2-3B-Instruct**, macro-averaged over SQuAD v2, LongBench-v2, LooGLE, and QASPER.



(a) Relative accuracy. (b) Relative completion speed. (c) Relative peak memory.

Figure 24: **LLMLingua-2 (learned, out-of-scope reference) on Qwen2.5-7B-Instruct**, macro-averaged over SQuAD v2, LongBench-v2, LooGLE, and QASPER.



(a) Relative accuracy. (b) Relative completion speed. (c) Relative peak memory.

Figure 25: **LLMLingua-2 (learned, out-of-scope reference) on Qwen3.5-9B**, macro-averaged over SQuAD v2, LongBench-v2, LooGLE, and QASPER.

Table 41: **LLMLingua-2 reference compression on Llama-3.2-3B-Instruct.** Task metrics, warm-cache completion time (t_{TC}), and pruning ratio (ρ) are reported.

Category	Method	LongBench v2			SQuAD v2			
		$Q_{abs}\uparrow$	$t_{TC}\downarrow$	ρ	EM ↑	F1 ↑	$t_{TC}\downarrow$	ρ
	Baseline	22.27	10.133	–	9.25	41.16	0.225	–
LLMLingua-2	LLMLingua2-20pct	24.25	9.208	18.66	9.44	40.09	0.226	18.94
	LLMLingua2-40pct	24.65	6.936	39.36	8.89	36.46	0.216	40.20
	LLMLingua2-60pct	25.25	5.034	59.42	9.64	31.97	0.210	58.91
	LLMLingua2-80pct	24.25	2.782	79.98	12.14	26.89	0.225	77.99

Category	Method	LooGLE			QASPER			
		ROUGE-L ↑	$t_{TC}\downarrow$	ρ	EM ↑	F1 ↑	$t_{TC}\downarrow$	ρ
	Baseline	11.24	3.009	–	4.48	33.71	0.755	–
LLMLingua-2	LLMLingua2-20pct	11.39	2.275	19.01	3.24	33.20	0.622	17.79
	LLMLingua2-40pct	11.08	1.822	39.42	3.03	31.43	0.584	38.85
	LLMLingua2-60pct	10.75	1.453	59.69	2.83	29.07	0.527	58.82
	LLMLingua2-80pct	9.78	1.125	79.51	3.65	26.85	0.430	79.03

Table 42: **LLMLingua-2 reference compression on Qwen2.5-7B-Instruct.** Task metrics, warm-cache completion time (t_{TC}), and pruning ratio (ρ) are reported.

Category	Method	LongBench v2			SQuAD v2			
		$Q_{abs}\uparrow$	$t_{TC}\downarrow$	ρ	EM ↑	F1 ↑	$t_{TC}\downarrow$	ρ
	Baseline	30.22	4.156	–	38.01	72.21	0.522	–
LLMLingua-2	LLMLingua2-20pct	33.00	4.551	17.97	36.20	69.80	0.512	18.48
	LLMLingua2-40pct	28.43	3.542	38.47	35.16	65.08	0.483	39.25
	LLMLingua2-60pct	30.02	3.001	58.60	35.90	59.16	0.440	57.72
	LLMLingua2-80pct	29.62	2.085	79.56	41.76	52.65	0.384	76.96

Category	Method	LooGLE			QASPER			
		ROUGE-L ↑	$t_{TC}\downarrow$	ρ	EM ↑	F1 ↑	$t_{TC}\downarrow$	ρ
	Baseline	10.81	8.395	–	8.68	43.21	1.960	–
LLMLingua-2	LLMLingua2-20pct	10.53	8.150	18.52	8.41	44.47	1.747	17.47
	LLMLingua2-40pct	10.69	7.441	38.58	8.13	43.04	1.548	38.37
	LLMLingua2-60pct	11.57	6.226	58.94	8.13	40.16	1.399	58.35
	LLMLingua2-80pct	12.08	5.174	79.15	9.44	35.13	1.153	78.63

Table 43: **LLMLingua-2 reference compression on Qwen3.5-9B.** Task metrics, warm-cache completion time (t_{TC}), and pruning ratio (ρ) are reported.

Category	Method	LongBench v2			SQuAD v2			
		$Q_{abs}\uparrow$	$t_{TC}\downarrow$	ρ	EM ↑	F1 ↑	$t_{TC}\downarrow$	ρ
	Baseline	35.79	4.853	–	35.67	70.04	0.286	–
LLMLingua-2	LLMLingua2-20pct	37.18	4.945	17.70	35.38	69.55	0.267	18.57
	LLMLingua2-40pct	36.78	3.948	38.27	35.37	66.29	0.242	39.41
	LLMLingua2-60pct	35.98	3.407	58.12	38.42	61.69	0.212	57.89
	LLMLingua2-80pct	30.62	2.455	78.38	44.57	54.53	0.187	77.08

Category	Method	LooGLE			QASPER			
		ROUGE-L ↑	$t_{TC}\downarrow$	ρ	EM ↑	F1 ↑	$t_{TC}\downarrow$	ρ
	Baseline	12.09	6.307	–	6.34	41.39	1.584	–
LLMLingua-2	LLMLingua2-20pct	11.98	5.936	17.56	6.00	42.11	1.427	17.46
	LLMLingua2-40pct	11.96	5.199	37.54	6.27	41.97	1.211	38.14
	LLMLingua2-60pct	10.71	4.431	57.49	7.31	40.72	1.007	57.88
	LLMLingua2-80pct	8.73	3.841	77.28	8.89	37.37	0.710	77.99

Table 44: **Qualitative comparison of tokenization patterns** for the model families evaluated in this study.

Model Family	Tokenizer	Segmentation Unit	Examples
LLaMA-3	Tiktoken (BPE)	Subword (Byte fallback)	un | belie | vably ther | m | odynam | ically state | -of | -the | -art Res | Net | 50
Mistral	SentencePiece (BPE)	Subword	un | bel | iev | ably therm | od | ynam | ically state | - | of | - | the | - | art Res | Net | 5 | 0
Qwen2.5	Byte-level BPE (BBPE)	Byte / Subword	un | belie | vably ther | m | odynam | ically state | -of | -the | -art Res | Net | 5 | 0
Qwen3	Byte-level BPE (BBPE)	Byte / Subword	un | believ | ably ther | m | odynam | ically state | -of | -the | -art Res | Net | 5 | 0
Qwen3.5	Byte-level BPE (BBPE)	Byte / Subword	un | believ | ably ther | m | odynam | ically state | -of | -the | -art Res | Net | 5 | 0
Phi-4	Tiktoken-compatible (BPE)	Subword (Byte fallback)	un | belie | vably ther | m | odynam | ically state | -of | -the | -art Res | Net | 50

I TOKENIZER-SPECIFIC SUBWORD STRUCTURE

Table 44 compares the subword segmentation of the tokenizers used by the evaluated model families: tiktoken-style BPE (Llama-3, Phi-4), SentencePiece BPE (Mistral), and byte-level BPE (Qwen2.5, Qwen3, Qwen3.5).

Segmentation diversity. Segmentation granularity differs between families. Mistral produces finer-grained splits (e.g., `un|bel|iev|ably`, and single digits for numbers), whereas Llama-3 and Phi-4 merge more aggressively (e.g., `un|belie|vably`, `50`). All tokenizers preserve the order of the input and respect whitespace-delimited word boundaries in these examples.

Consequences for word-level interventions. The lexical, frequency, and syntactic rules (L-, F-, and D-series) operate on whole words before tokenization, and dependency roles such as subject, predicate, and object are properties of words rather than tokens. Their selection is therefore identical for all models; only the achieved token-level reduction ρ depends on the tokenizer. The subword rules S1 and M1 are more tokenizer-sensitive in their achieved reduction: S1, for example, removes 23.4% of tokens for the Llama models, 24.6% for Qwen2.5, 25.0% for Qwen3.5, and 28.5% for Mistral (Tab. 11), because finer-grained tokenizers spend more tokens on the removed fragments.

Implications for generalization. Despite these differences, intervention rankings are highly correlated across models with different tokenizers (App. C.4), and D2 has the largest reduction-adjusted quality among the linguistic rules for models of all three tokenization schemes. This indicates that the sensitivities we characterize are not artifacts of a particular vocabulary or merging heuristic. It does not imply that they are universal: the multilingual results (App. G) show that the same word-level rules can behave very differently when the linguistic structure of the input changes.