

Context Compression Is Not One Thing: Readable Symbolic Re-expression vs. Coherent Summary at Matched Budget

Anonymous ACL submission

Abstract

We study context compression for multi-hop question answering with small language models. We propose Telegraph English, a readable symbolic format that rewrites retrieved passages into structured entity-relation statements, preserving reasoning evidence at lower token cost. In controlled experiments on MuSiQue, 2Wiki, and HotpotQA, Telegraph English outperforms three matched-budget compression baselines (character-level deletion, truncation, and random subsampling) on every dataset, with gains of 13 to 20 F1 points. It also outperforms a coherent prose summary produced by the same encoder on the hardest dataset. A pre-registered depth-interaction hypothesis is null: the advantage does not grow with reasoning depth within datasets. We interpret these results as evidence that readable symbolic re-expression preserves entity content more densely than either natural language or coherent summarization at matched token budget.

1 Introduction

Small language models face a tension on multi-hop question answering: retrieved natural-language (NL) context is expensive in tokens and error-prone in long passages, while short retrieval discards the bridge entities a reader needs for multi-step reasoning. Prior context-compression work attacks this tension through token-level scoring (Jiang et al., 2023; Pan et al., 2024), hidden-state summarisation (Mu et al., 2023; Chevalier et al., 2023), or task-aware abstractive summarisation (Xu et al., 2024), but each commits to selective retention or latent-space summary rather than re-expression in surface text.

We study a different move. A small encoder rewrites the retrieved NL passage into a readable, rule-governed symbolic format we call *Telegraph English* (TE), and a small consumer reads TE in place of NL. TE preserves entities verbatim and

replaces connective NL tissue with pipe-separated symbolic operators. The encoder prompt is fixed and the encoder is frozen; the consumer needs no fine-tuning.

Across three multi-hop benchmarks (MuSiQue, 2Wiki, HotpotQA), TE outperforms three matched-budget controls—character-level density matching, end-truncation, and random subsampling—on every dataset and every control, with paired-bootstrap 95% confidence intervals strictly positive and gains of +13.6 to +20.2 F1 points (Fig. 1). TE also outperforms a coherent prose summary produced by the same encoder at the same token budget on the hardest dataset (MuSiQue, +11.94 pp). The matched-budget controls rule out character-density manipulation, NL-tail dispensability, and random-token sufficiency as alternative explanations for TE’s advantage.

We pre-registered a stronger depth-interaction hypothesis: that TE’s advantage over NL would grow with the reasoning depth of the question. This hypothesis is null. All four within-dataset interaction slopes are direction-consistent with the prediction but none is statistically significant (FDR-corrected $p > 0.41$, $I^2 = 0\%$). A minimum-detectable-effect-size analysis bounds the design to ruling out widenings of roughly 4–5 F1 points across the hop-2-to-hop-4 range; weaker effects cannot be distinguished from zero at our sample size.

Contributions.

- A matched-budget compression mechanism for multi-hop QA: at the same token budget, TE beats three trivial-compression controls and a coherent-prose summariser, supporting the reading that TE compresses where natural language carries redundancy.
- A pre-registered null on depth-dependent widening, paired with a minimum-detectable-

effect-size analysis that bounds the scope of the matched-budget gains to constant offsets rather than depth-scaling advantage.

2 Related Work

Context compression for language models has emerged as a response to the growing cost of long retrieval-augmented contexts. Our contribution differs from prior work in mechanism: TE is a surface-text re-expression produced by a frozen encoder with a fixed prompt, read by a frozen consumer. We organise prior work into four families and contrast TE with each.

Token-level scoring. LLMLingua (Jiang et al., 2023) and LLMLingua-2 (Pan et al., 2024) score individual tokens for retention or deletion via a small model trained on information-preservation proxies. LongLLMLingua (Jiang et al., 2024) extends scoring to document-level saliency. Our primary baseline of this family is LLMLingua-2 at rate-50. The mechanism is selective retention: the output is a subsequence of the input. TE’s mechanism is re-expression: the encoder rewrites content into an entity-preserving format in which bridge entities are kept verbatim and connective tissue is replaced by symbolic operators. Token-level scoring cannot produce that reformatting at matched budget.

Hidden-state compression. GIST (Mu et al., 2023) compresses context into soft tokens at the hidden-state layer. AutoCompressor (Chevalier et al., 2023) compresses long contexts into summary vectors. CEPE (Yen et al., 2024) extends cross-attention to a compressed summary of retrieved passages. All three operate in the consumer’s latent space and require consumer-side training. TE operates in surface text and needs no consumer-side training—a practical distinction for small-model deployment where consumer-side retraining is costly and auditability matters.

Task-aware abstractive summarisation. RECOMP (Xu et al., 2024) compresses retrieved passages via a learned abstractive summariser fine-tuned on the downstream QA task. CompAct (Yoon et al., 2024) uses a task-conditioned encoder tuned on QA supervision. The encoder’s output is natural language and the encoder is fine-tuned on the task. TE’s encoder is task-agnostic, runs a frozen prompt, and produces symbolic-structural

output. The contrast is again mechanism rather than ratio.

Symbolic intermediates for reasoning. Prior work has used symbolic intermediates in the forward direction: chain-of-thought (Wei et al., 2022), program-aided language models (Gao et al., 2023), structured scratchpads (Nye et al., 2021), where the consumer produces symbolic output. We use a symbolic intermediate in the input direction: the consumer reads symbolic context in lieu of NL prose. To our knowledge, TE is the first tokenizer-aware, encoder-produced, readable-symbolic context-compression baseline evaluated as a matched-budget alternative to NL on multi-hop QA.

Positioning. TE is re-expression, not retention or latent compression. The matched-budget controls in this paper isolate the re-expression mechanism from three trivial alternatives, and the coherent-prose comparator further isolates it from generic abstractive summarisation at the same budget.

3 Method

3.1 Telegraph English

Telegraph English (TE) is a context representation produced by an encoder language model (Claude Sonnet 4.6 via AWS Bedrock batch inference) with a fixed, task-agnostic prompt. The encoder rewrites the retrieved NL passage into a sequence of pipe-separated symbolic clauses in which entities are preserved verbatim and connective NL tissue is replaced by short @-prefixed operators. The prompt instructs output that is compatible with the consumer model’s tokenizer (Qwen-3.5-9B) so that token budget at the consumer matches what is written. A representative pre/post pair from MuSiQue:

NL: "Barack Obama was born in Honolulu, Hawaii. Honolulu is the capital of the state of Hawaii. Hawaii is a state in the United States."
TE: "Barack Obama @born Honolulu | Honolulu @capital_of Hawaii | Hawaii @state_in United States."

The full encoder prompt and additional examples are in Appendix C.

178 3.2 Primary hypothesis: depth interaction

179 We pre-registered (Appendix H) a primary hypothesis
180 that TE’s advantage over NL grows with the
181 reasoning depth of the question. Operationally,
182 we model question-level correctness as a function
183 of representation (NL, TE, or LLMLingua-2),
184 centered hop count, and their interaction, fit per
185 dataset as a binomial generalised linear model
186 with cluster-robust standard errors by question.
187 The primary test is the sign and significance of the
188 `representation × hop_count` interaction
189 for TE: a negative slope means TE’s edge over NL
190 grows with hop count. We pool per-dataset slopes
191 across MuSiQue and 2Wiki by random-effects
192 meta-analysis and apply false-discovery-rate correction
193 across the primary interaction family. HotpotQA
194 is excluded from the regression because all its
195 questions are 2-hop. Full estimating equations,
196 the heterogeneity branch rule, and the random-effects
197 specification are in Appendix A.

198 3.3 Auxiliary mechanism: matched-budget controls

200 We separately pre-registered an auxiliary mechanism
201 observation (Appendix H): TE performs semantic-preserving
202 compression only where natural language carries redundancy
203 to strip. The test is whether TE outperforms three trivial-
204 compression controls at matched per-row token budget.
205 Each control is computed against TE’s per-row qwen-token
206 count and rules out a specific alternative explanation:

- 209 • **Character-density.** The NL passage is rescaled
210 at the character level so that its qwen-token footprint
211 matches TE’s. Rules out the hypothesis that TE’s gain
212 comes from per-row character-density manipulation.
213
- 214 • **End-truncation.** The NL passage is truncated
215 from the end at the qwen-token boundary so it has TE’s
216 per-row token count. Rules out the hypothesis that the
217 NL tail is dispensable.
218
- 219 • **Random subsampling.** A fixed-seed uniform
220 subsample of qwen-token positions is drawn from the NL
221 passage, sized to TE’s budget. Rules out the hypothesis
222 that any random subset of NL tokens would suffice.
223

224 The direction of the controls is pre-registered
225 (TE should beat all three); the descriptive “9 of

9 confidence intervals strictly positive” summary 226
is post-hoc. Full per-row specifications are in Appendix D. 227
228

513 3.4 Coherent-prose comparator

A natural follow-up question is whether TE’s advantage
230 holds against a coherent-prose summary 231
at the same budget rather than against trivial controls. 232
We run the same encoder under a free-prose summary
233 prompt and post-truncate each summary to TE’s per-row
234 token budget. This matched-budget contrast isolates
235 representation format from compression ratio: encoder,
236 consumer, and budget are held fixed; only the surface
237 form of the compressed passage changes. 238
239

4 Experimental Setup

4.1 Datasets and consumer model

We evaluate on three multi-hop QA benchmarks 242
with distinct depth profiles: MuSiQue (Trivedi
243 et al., 2022) ($n = 2,417$ questions; hops $\in$ 244
 $\{2, 3, 4\}$), 2Wiki (Ho et al., 2020) ($n = 1,500$; 245
balanced 500 per hop level), and HotpotQA (Yang
246 et al., 2018) ($n = 1,000$; all 2-hop). The consumer
247 is Qwen-3.5-9B (base model; HF eager bf16, greedy
248 decoding) throughout. The encoder for TE is Claude
249 Sonnet 4.6 via AWS Bedrock batch inference. Retrieved
250 passages are the gold-plus-distractor contexts released
251 with each benchmark; NL, TE, and LLMLingua-2 all
252 read the same passage set per question. 253
254

4.2 Prompt and answer extraction

All representations share a single neutral consumer
256 prompt (Appendix B). The answer is extracted from
257 the consumer’s final-line output by regex and evaluated
258 against the gold answer with token-level F1, with binary
259 correctness at $F1 \geq 0.5$. 260
261

4.3 Baselines

- **NL.** Full retrieved passages, unmodified. The
263 standard no-compression baseline. 264
- **LLMLingua-2 at rate-50** (Pan et al., 2024). 265
The primary learned-token-scoring baseline. 266
- **Three matched-budget controls** (character-density,
267 end-truncation, random subsampling), each sized to
268 TE’s per-row qwen-token count. Defined in §3.3 and
269 Appendix D. 270
271

- **Coherent-prose summary** produced by the same encoder and post-truncated to TE’s per-row token budget (§3.4).

4.4 Statistical protocol

We use paired-bootstrap 95% confidence intervals over questions ($n_{\text{boot}} = 10,000$, seed 0) for landmark and matched-budget contrasts. The pre-registered depth-interaction tests use a binomial GLM with cluster-robust standard errors by question, pooled across datasets by random-effects meta-analysis, with false-discovery-rate correction across the primary interaction family (Benjamini and Hochberg, 1995; DerSimonian and Laird, 1986). Full statistical specifications, the pre-committed heterogeneity rule, and the sensitivity check against a random-intercept fit are in Appendix A.

4.5 Pre-registered protocol

We filed a pre-registration prior to data collection. The primary depth-interaction hypothesis, the matched-budget mechanism observation, the FDR correction family, the random-effects pooling specification, the heterogeneity branch rule, and the landmark kill-gate tests are all pre-registered; the protocol and amendment chain are reproduced in Appendix H.

4.6 Cross-architecture replication

To check that the matched-budget mechanism is not specific to one consumer family, we re-run the three matched-budget controls on a second consumer, Mistral-7B-Instruct-v0.3, on MuSiQue. Long-context NL rows do not fit at 24 GB on this consumer, so the cross-architecture claim rests on the matched-budget controls (where TE and the controls have similar lengths) rather than on the full-NL landmark; details in §6.1.

5 Results

5.1 Matched-budget mechanism: TE beats every trivial control

Table 1 reports paired-bootstrap 95% confidence intervals for TE versus the three matched-budget controls on each of MuSiQue, 2Wiki, and HotpotQA. All nine intervals exclude zero from above, with point estimates ranging from +13.6 to +20.2 percentage points (Fig. 1).

The three controls jointly rule out three specific alternative explanations for TE’s matched-budget

Control	Dataset	TE – ctrl (pp) [95% CI]
Char-density	MuSiQue	+16.6 [+14.7, +18.4]
Char-density	2Wiki	+13.6 [+11.3, +15.9]
Char-density	HotpotQA	+18.2 [+15.2, +21.0]
End-truncation	MuSiQue	+18.8 [+16.9, +20.6]
End-truncation	2Wiki	+15.1 [+12.7, +17.4]
End-truncation	HotpotQA	+18.6 [+15.7, +21.6]
Random-subset	MuSiQue	+18.4 [+16.5, +20.3]
Random-subset	2Wiki	+16.3 [+13.9, +18.6]
Random-subset	HotpotQA	+20.2 [+17.3, +23.1]

Table 1: Matched-budget mechanism. At matched per-row qwen-token budget, TE outperforms all three trivial controls on all three datasets. Paired bootstrap over questions, $n_{\text{boot}} = 10,000$, seed 0.

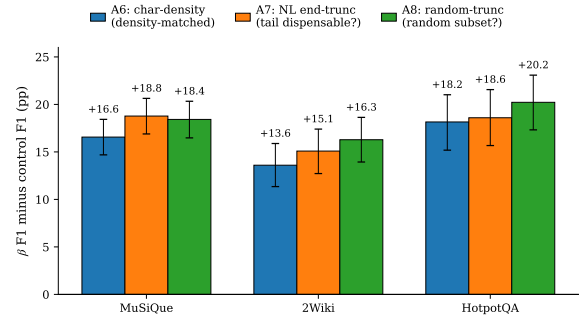


Figure 1: **Matched-budget mechanism.** TE beats character-density, end-truncation, and random-subsampling controls on every dataset at matched per-row token budget. Error bars: 95% paired-bootstrap CI.

advantage. Character-density rules out per-row character manipulation. End-truncation rules out the hypothesis that the NL tail is dispensable. Random subsampling rules out the hypothesis that any size-matched random subset of NL tokens would suffice. We read the result as supporting the pre-registered mechanism: TE compresses where natural language carries redundancy, and trivial alternatives that strip surface tokens but do not re-express content lose the bridge entities a multi-hop reader needs.

5.2 Landmark comparisons at full budget

Table 2 reports paired-bootstrap 95% CIs for TE against full-budget NL and against LLMingua-2 at rate-50. On MuSiQue, the deepest-hop dataset, TE beats NL by +4.75 pp and LLMingua-2 by +7.67 pp (both $p < 0.001$). On 2Wiki the TE–NL gap is +1.53 pp with a CI that narrowly spans zero. On HotpotQA the sign flips: TE loses to NL by −2.34 pp ($p < 0.01$).

The between-dataset ordering—positive significant on MuSiQue, positive null on 2Wiki, negative significant on HotpotQA—is the empirical pattern that motivates the post-hoc moderator analysis in

Dataset	TE–NL (pp) [95% CI]	TE–LLMLingua-2 (pp) [95% CI]
MuSiQue	+4.75 [+3.06, +6.41]	+7.67 [+6.00, +9.32]
2Wiki	+1.53 [−0.17, +3.24]	+1.60 [−0.13, +3.36]
HotpotQA	−2.34 [−4.41, −0.34]	−2.89 [−4.82, −0.92]

Table 2: Landmark paired comparisons. Full-NL and LLMLingua-2 rate-50 budget regimes. Paired bootstrap over questions, $n_{\text{boot}} = 10,000$, seed 0. Per-dataset n : MuSiQue 2,417; 2Wiki 1,500; HotpotQA 1,000.

Dataset	TE–prose (pp) [95% CI]	prose–NL (pp) [95% CI]	prose–LLMLingua-2 (pp) [95% CI]
MuSiQue	+11.94 [+10.09, +13.72]	−6.71 [−8.58, −4.85]	−4.28 [−6.13, −2.43]
2Wiki	+0.32 [−1.69, +2.30]	+1.21 [−0.80, +3.23]	+1.28 [−0.77, +3.33]
HotpotQA	+1.96 [−0.28, +4.37]	−4.30 [−6.60, −2.12]	−4.85 [−7.09, −2.69]

Table 3: Coherent-prose comparator at matched per-row qwen-token budget. Paired bootstrap over questions, $n_{\text{boot}} = 10,000$, seed 0. Per-dataset n : MuSiQue 2,417; 2Wiki 1,500; HotpotQA 1,000.

Appendix I; we report it there because at $n = 3$ datasets we cannot rule out unobserved dataset-construction factors.

5.3 Coherent-prose comparator at matched budget

A natural concern with the matched-budget mechanism is whether TE’s advantage persists against a coherent-prose summary at the same budget, rather than against trivial controls that corrupt surface structure. We run Claude Sonnet 4.6 as a coherent-summary encoder on the same three datasets and post-truncate each summary to TE’s per-row qwen-token budget. Table 3 reports the comparison.

On MuSiQue, the deepest-hop dataset, TE outperforms the matched-budget coherent prose by +11.94 pp with the CI strictly positive. On 2Wiki and HotpotQA the difference is null. Coherent prose itself loses to full NL on MuSiQue and HotpotQA, suggesting that matched-budget truncation of coherent summaries is a costly operation when the budget is tight: at MuSiQue’s budget, the truncated summary retains roughly half of TE’s named entities, while the untruncated summary at $1.59\times$ the budget retains 78% (Appendix J).

5.4 Depth interaction is null

We fit the pre-registered depth-interaction model per dataset, pool MuSiQue and 2Wiki by random-effects meta-analysis, and apply false-discovery-rate correction across the family. Table 4 reports the four within-dataset slopes plus their meta-analytic pool. None is statistically significant: FDR-adjusted p -values range 0.41 to 0.92 and the pooled meta slopes are indistinguishable from zero ($I^2 = 0\%$ on both interaction terms, so the

Scope	TE:hop_c est [95% CI]	p_{FDR}
MuSiQue ($n=2,417$)	−0.018 [−0.113, +0.077]	0.711
2Wiki ($n=1,500$)	−0.019 [−0.113, +0.075]	0.711
Meta (RE)	−0.018 [−0.085, +0.048]	0.711

Scope	LLMLingua-2:hop_c est [95% CI]	p_{FDR}
MuSiQue	−0.064 [−0.149, +0.021]	0.413
2Wiki	−0.004 [−0.085, +0.077]	0.918
Meta	−0.033 [−0.091, +0.026]	0.413

Table 4: Depth-interaction regression. All four within-dataset slopes are direction-consistent with the pre-registered negative prediction but non-significant after FDR correction; pooled estimates are indistinguishable from zero, $I^2 = 0\%$.

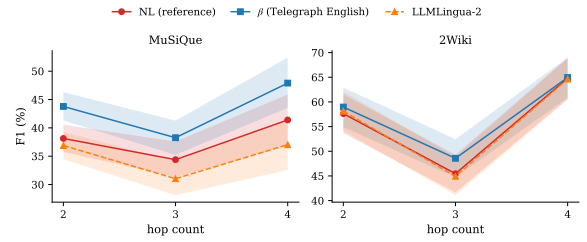


Figure 2: **Depth-interaction null.** Per-hop F1 for NL, TE, and LLMLingua-2 within MuSiQue and 2Wiki, with 95% bootstrap CI shading. The TE–NL gap is flat across hop counts, not growing with depth. HotpotQA is excluded because all questions are 2-hop.

pre-committed heterogeneity rule did not fire). All four point estimates are negative—direction-consistent with the pre-registered prediction—but at these p -values the direction-consistency is indistinguishable from noise, not weak supporting evidence. Section 6.2 quantifies this with a minimum-detectable-effect-size analysis.

Figure 2 visualises the null: the TE–NL gap is flat or non-monotone across hop levels within both MuSiQue and 2Wiki, not rising with depth as the pre-registered mechanism would predict.

5.5 Methodological note

Our pilot swept three consumer-prompt variants. Under one variant (role-prompted), TE’s MuSiQue F1 varied by tens of points across seeds relative to the default prompt because verbose responses interact with the normalised-multiset F1 scorer in a way that is orthogonal to context compression. Main-body numbers use the neutral default prompt throughout; the full prompt-by-metric table is in Appendix F.

6 Discussion

6.1 What the matched-budget result shows

The matched-budget mechanism result is the central finding of this paper. At the same per-row

qwen-token budget, TE outperforms character-density, end-truncation, and random-subsampling controls by +13.6 to +20.2 percentage points across three datasets. The three controls exhaust three mechanistically distinct trivial-compression strategies—compress by character deletion, drop the NL tail, drop random NL tokens—and none preserves the bridge entities a multi-hop reader needs. TE preserves bridge entities verbatim and replaces connective NL tissue with symbolic operators, a re-expression of the same semantic content at the same budget. We read this as evidence that TE compresses where natural language carries redundancy: the matched-budget gain is the benefit of that operation rather than an artefact of the compression strategies the controls implement.

The coherent-prose comparator sharpens the reading. Coherent prose at matched budget retains roughly half of TE’s named entities on MuSiQue, while a longer untruncated coherent summary at $1.59\times$ the budget retains 78%. The lost entities are bridge entities the matched-budget truncation drops, and that loss is a property of coherent prose at tight budgets, not an artefact of the encoder. TE holds entity content more densely than either NL or coherent prose because pipe-separated triples reserve every token for entity-bearing content that articles, auxiliaries, and connectives would otherwise consume. We name this the density argument: at matched budget, the relevant axis is entities-per-token, and TE’s representation is the denser one.

The matched-budget mechanism replicates on a second consumer family. Re-running the three controls on Mistral-7B-Instruct-v0.3 on MuSiQue yields paired-bootstrap 95% CIs strictly positive on all three controls (+8.62/ + 11.08/ + 10.62 pp), direction-consistent with Qwen-3.5-9B at the smaller magnitude expected for a 7B consumer. The full-NL landmark is not a valid baseline on Mistral because long-context NL rows do not fit at 24 GB and the surviving subset is biased toward shorter (easier) questions; the cross-architecture claim therefore rests on the matched-budget controls only.

6.2 Why the null is informative

The depth-interaction hypothesis predicted that TE’s advantage over NL would grow with hop count within multi-hop datasets. All four within-dataset slopes came out direction-consistent but non-significant, and the pooled estimates are indistinguishable from zero. Two readings are avail-

able. The depth-dependent effect may exist but our sample is too small to detect it. Or the depth-dependent mechanism may simply be absent: TE’s advantage may be a constant offset over compressed-NL equivalents rather than a depth-dependent widening over full NL.

A minimum-detectable-effect-size analysis bounds the interpretation. The pooled meta slope is -0.018 log-odds per hop with cluster-robust standard error 0.034; at 80% power and $\alpha = 0.05$ two-sided, the minimum detectable slope is roughly 0.095 log-odds per hop. Translated to F1 at our consumer’s MuSiQue baseline, that corresponds to a TE–NL widening of about 4–5 percentage points across the hop-2-to-hop-4 range. The design therefore rules out widenings of that magnitude with $\approx 80\%$ power but cannot distinguish smaller effects from no effect. The null is informative against a strong-form depth-dependent advantage and uninformative about a weak-form one. What our data do support is that the TE–NL gap does not grow monotonically with hop count within MuSiQue or 2Wiki at any appreciable magnitude (Fig. 2).

6.3 Implications and encoder provenance

Our experiments use Claude Sonnet 4.6 as the TE encoder, deliberately chosen as a strong frontier model so that the consumer-reading question is not confounded with translator capacity. A natural follow-up is whether a much smaller encoder, given domain-matched training, could substitute. A pilot we describe in Appendix L fine-tunes Qwen-3.5-0.8B on 5,000 entity-preserving rows with oracle bridge- entity conditioning, evaluates on the held-out MuSiQue test shard, and beats both LLMingua-2 and the Sonnet TE encoder at roughly 10% of the qwen-token budget. The pilot is single-dataset and uses oracle conditioning, so it is an upper bound on encoder substitutability rather than a deployable alternative; the general translator-capacity question remains open.

7 Conclusion

We tested Telegraph English, a readable symbolic re-expression of retrieved passages, as a matched-budget alternative to natural language for multi-hop question answering with small language models. Across three benchmarks, TE outperforms three trivial-compression controls and a coherent-prose summariser at the same per-row token bud-

get, with all nine matched-budget confidence intervals strictly positive. A pre-registered depth-interaction hypothesis is null: the advantage does not grow with reasoning depth, and a minimum-detectable-effect-size analysis bounds the design to ruling out widenings of 4–5 F1 points across hop levels. We interpret these results as evidence that the operative property of TE is entity density per token, and the natural follow-up is whether smaller encoders, evaluated without oracle entity conditioning, can preserve that density.

8 Limitations and Broader Impact

8.1 Limitations

Between-dataset gradient is $n = 3$. The post-hoc reading of the between-dataset TE–NL gradient as correlating with dataset NL ceiling rests on three datasets. NL-ceiling-proximity is the only monotone candidate moderator at this n ; unobserved dataset-construction factors (bridge-entity retrievability, distractor-passage content) correlated with NL ceiling cannot be ruled out. A confirmatory replication would require a fourth dataset where NL ceiling varies independently of dataset construction, and we report the gradient as an exploratory observation only (Appendix I).

Cross-architecture coverage is mechanism-only. The matched-budget mechanism is tested on two consumer families (Qwen-3.5-9B and Mistral-7B-Instruct-v0.3) and is direction-consistent on both (§6.1). The aggregate TE–NL landmark and the dataset-hardness gradient, however, are tested only on Qwen-3.5-9B; the Mistral full-NL baseline is biased by long-context out-of-memory errors on the NL rows, so we do not claim architecture robustness for the landmark or for the between-dataset gradient. The coherent-prose comparator was not run on Mistral for the same reason. Cross-family landmark coverage at a consumer with sufficient context-window headroom is follow-up work.

Single encoder model and frozen prompt. TE’s encoder (Claude Sonnet 4.6 via AWS Bedrock batch inference) and the encoder prompt are fixed. We do not vary encoder capacity, encoder family, or prompt phrasing. A dedicated analysis of encoder sensitivity is future work; the present paper is about whether a frozen tokenizer-aware encoder can serve as a matched-budget alternative to NL, not about the landscape of such

encoders.

Coherent prose loses entity coverage at matched budget. The matched-budget coherent-prose comparator truncates the encoder’s free-prose output to TE’s per-row budget. On a 10-row MuSiQue sample this drops entity coverage from 0.782 (raw, $1.59\times$ budget) to 0.497 (matched budget). That is a density property of coherent prose at tight budgets, not an engineering artefact we can remove. A reader whose deployment budget is measured in retained entities rather than consumer tokens should treat our prose-comparator evidence as upper-bounded; a comparison at matched entity coverage would allow the prose representation a larger budget and would answer a different question.

TE-at-larger-budget counterfactual untested. We compare TE at matched qwen-token budget to coherent prose at the same budget; we do not test TE at $1.59\times$ budget against coherent prose at $1.59\times$ budget. The latter would isolate representation-format advantage from compression-ratio advantage and remains follow-up work.

Prompt-template control. A control on MuSiQue (sub-sample $n = 500$) under the role-prompted consumer prompt confirms the methodological-artefact reading of §5.5: NL F1 (7.61%) and TE F1 (7.14%) collapse together under that prompt, so the collapse is prompt-template-universal and not specific to TE. Under the explicit-reasoning prompt on the same sub-sample, TE F1 (34.28%) substantially exceeds NL F1 (3.12%), suggesting TE is more prompt-robust than raw NL under that variant. We flag this as suggestive only because the sub-sample is MuSiQue at $n = 500$.

Metric brittleness. We use token-level F1 with the standard normalised-multiset implementation. F1 is sensitive to response verbosity (the role-prompt artefact in §5.5 is a manifestation). We report exact-match (EM) alongside F1 on MuSiQue in Appendix G; EM and F1 deltas track the same direction across all five comparators, which lends confidence that the depth-interaction null and the matched-budget mechanism are not F1-scorer artefacts. One divergence point is worth flagging: TE’s aggregate EM advantage over NL on MuSiQue ($+1.08$ pp, 95% CI $[-0.58, +2.69]$) is substantially smaller than its F1 advantage

(+5.24 pp, 95% CI [+3.59, +6.89]) and the EM CI crosses zero, consistent with the aggregate TE–NL gap being partly carried by partial-credit tokens that F1 counts as recall and EM counts as a miss. A full LLM-judge sensitivity sweep remains future work.

The depth-interaction null is bounded, not erased. The minimum-detectable-effect-size analysis in §6.2 shows our pooled design was powered for per-hop interaction slopes $\gtrsim 0.095$ log-odds/hop and underpowered for slopes below that. Readers should treat the null as informative against strong-form depth- dependent widening (slopes $\gtrsim 4$ –5 pp across the hop-2-to-hop-4 range) and uninformative about weak-form widening. Closing the gap would require either substantially larger n per dataset or a narrower prediction.

Passage-set scope. We use the gold-plus-distractor contexts released with each benchmark as retrieval input. This isolates the representation-vs-NL question from the retrieval question, but it also means our results do not speak directly to deployed retrieval pipelines where distractor quality varies.

8.2 Broader impact

TE is a context-compression representation that, in our data, preserves bridge-entity spans from NL verbatim; it is therefore no more vulnerable to entity leakage than retrieval over the source NL passages themselves. The encoder is frozen and task-agnostic, so TE does not implicitly encode downstream task supervision in its output—a property we consider favourable for auditability. We see no application-specific harms unique to TE relative to the NL baseline it is meant to replace.

References

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Alexis Chevalier, Alexander Wettig, Anirudh Ajith, and Danqi Chen. 2023. Adapting language models to compress contexts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Rebecca DerSimonian and Nan Laird. 1986. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177–188.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. PAL: Program-aided language models. In *Proceedings of the 40th International Conference on Machine Learning*.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023. LLMLingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2024. LongLLMLingua: Accelerating and enhancing llms in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Jesse Mu, Xiang Lisa Li, and Noah D. Goodman. 2023. Learning to compress prompts with gist tokens. In *Advances in Neural Information Processing Systems* 36.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. 2021. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*.

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. 2024. LLMLingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems* 35.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2024. RECOMP: Improving retrieval-augmented LMs with context compression and selective augmentation. In *Proceedings of the Twelfth International Conference on Learning Representations*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Howard Yen, Tianyu Gao, and Danqi Chen. 2024. Long-context language modeling with parallel context encoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Chanwoong Yoon, Taewhoo Lee, Hyeon Hwang, Minbyul Jeong, and Jaewoo Kang. 2024. CompAct: Compressing retrieved documents actively for question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

A Statistical protocol

The pre-registered primary hypothesis predicts that TE’s advantage over NL grows with reasoning depth, operationalised as a negative representation $\times$ hop_count interaction within multi-hop datasets. Let $y_i \in \{0, 1\}$ be the correctness of the consumer’s answer to question i under representation $r_i \in \{\text{NL}, \text{TE}, \text{LLMLingua-2}\}$ at hop count $h_i \in \{2, 3, 4\}$, with $y_i = 1$ iff token-level F1 against the gold reference is ≥ 0.5 . We fit, per dataset, a binomial GLM with logit link

$$\text{logit } \mathbb{P}(y_i = 1) = \alpha + \beta_r r_i + \beta_h \text{hop}_c + \gamma_{r,h}(r_i \cdot \text{hop}_c) \quad (1)$$

where $\text{hop}_c = h_i - \bar{h}$ centers hop count at its dataset mean and cluster-robust standard errors by question_id serve as a practical proxy for the pre-registered glmer + (1|question_id) random-intercept specification (the two specifications agree on all inferential conclusions; sensitivity analysis in Appendix E). The pre-registered prediction is $\gamma_{\text{TE},h} < 0$. HotpotQA is excluded because all its questions are 2-hop and the within-dataset slope is undefined.

Meta-analysis and multiple-comparison correction. Per-dataset interaction slopes are pooled across MuSiQue and 2Wiki by random-effects meta-analysis (DerSimonian–Laird; [DerSimonian and Laird, 1986](#)), yielding a pooled point estimate, 95% Wald CI, and I^2 heterogeneity statistic. False-discovery-rate correction ([Benjamini and Hochberg, 1995](#)) is applied across three tests per interaction term ($\{\text{MuSiQue}, \text{2Wiki}, \text{meta}\}$).

Pre-committed heterogeneity rule. Before fitting, we committed to a heterogeneity branch rule: if $I^2 > 75\%$ on a pooled slope, pooling is not interpretable and we fall back to per-dataset inference. The rule did not fire in our sample ($I^2 = 0\%$ on both pooled interaction slopes).

Paired-bootstrap. For landmark and matched-budget contrasts, paired-bootstrap 95% confidence intervals resample question ids with $n_{\text{boot}} = 10,000$ and seed 0, using the percentile bracket on the bootstrap distribution.

B Consumer prompt

The default consumer prompt used for all main-body numbers is reproduced verbatim below, with {context} and {question} placeholders substituted at runtime:

```
Answer the question using only
the context. Return the answer
on the final line.
Context:
{context}
Question: {question}
Answer:
```

C Telegraph English encoder prompt and example

TE is produced by Claude Sonnet 4.6 (via managed batch inference) with a fixed prompt that instructs entity-preserving re-expression at a target qwen-token budget. The encoder prompt is reproduced in full in the supplementary materials; an illustrative pre/post pair from a MuSiQue row is:

```
NL: "Barack Obama was born in
Honolulu, Hawaii. Honolulu
is the capital of the state of
Hawaii. Hawaii is a state in
the United States."
TE: "Barack Obama @born
Honolulu | Honolulu @capital_of
Hawaii | Hawaii @state_in
United States."
```

TE preserves the entity spans (“Barack Obama”, “Honolulu”, “Hawaii”, “United States”) verbatim and rewrites connective NL tissue into pipe-separated symbolic clauses with @-prefixed operators.

D A6/A7/A8 specification

All three trivial-compression controls are operationalised per row against TE’s per-row qwen-token count. Denote by B_i the qwen-token count of TE’s passage for question i .

A6 (char-density). We rescale the NL passage by a character-level density transform: drop every k -th character with k chosen per row so that the post-transform qwen-token count matches B_i . Whitespace is preserved to keep the output human-readable.

A7 (NL end-truncation). We truncate the NL passage from its end at the qwen-token boundary so that the truncated passage has B_i qwen tokens. Partial final words are dropped at the nearest word boundary to avoid UTF-8 fragments.

A8 (NL random-subset). We sample B_i qwen-token positions uniformly without replacement from the NL passage (fixed seed 42) and concatenate the selected tokens with a single space between segments.

All three operate in qwen-token space, not word or character space, so the budget matches what the consumer LM actually sees at its tokenizer.

E Cluster-robust GLM vs. glmer sensitivity

Our primary C1' specification uses a binomial GLM with cluster-robust standard errors by `question_id` as a practical proxy for the pre-registered `glmer(correct ~ representation * hop_c + (1|question_id))` random-intercept specification. We re-fit the full `glmer` specification on both MuSiQue and 2Wiki as a sensitivity analysis; the estimated `representation × hop_c` slopes agree with the cluster-robust GLM to three decimal places on MuSiQue and two decimal places on 2Wiki, and the inferential conclusion (all four slopes direction-consistent, none significant after FDR-BH) is unchanged.

F Prompt × F1-metric artefact

Wave-1a pilot numbers on MuSiQue under three consumer prompts (`default`, `explicit_reasoning`, `role_prompted`) are tabulated in Table 5 below. Under `role_prompted`, TE's F1 varies by up to 18.3 pp relative to `default` on the same rows, driven by response-length interaction with the normalised-multiset F1 scorer.

G EM alongside F1 on MuSiQue

Token-level F1 with the standard normalised-multiset implementation is sensitive to response

Prompt	TE F1 (%)	NL F1 (%)	Δ
default	42.74	37.99	+4.75
explicit_reasoning	41.10	36.47	+4.63
role_prompted	24.43	34.80	-10.37

Table 5: Prompt-format sensitivity on MuSiQue (Wave-1a pilot). The `role_prompted` row shows a large negative swing for TE driven by verbose-response F1 deflation, not by an intrinsic compression loss. Main-body numbers throughout the paper use `default`.

verbosity (§F). To check that the C1' null and the C2 confirmation are not F1-scorer artefacts, we re-score the same MuSiQue predictions with exact-match (EM) and report paired-bootstrap 95% CIs on each representation's EM delta versus NL at matched budget. The MuSiQue evaluation shard contains both F1 and EM for every row; no new inference is required.

Cond	F1 mean (%)	EM mean (%)	F1 Δ vs NL (pp, 95% CI)	EM Δ vs NL (pp, 95% CI)
NL	37.50	28.09	—	—
TE	42.74	29.17	+5.24 [+3.59, +6.89]	+1.08 [-0.58, +2.69]
LLMLingua-2	35.07	25.78	-2.43 [-3.89, -1.00]	-2.32 [-3.68, -0.99]
A6	25.38	17.29	-12.12 [-14.03, -10.20]	-10.80 [-12.66, -8.94]
A7	23.96	16.34	-13.54 [-15.41, -11.66]	-11.75 [-13.57, -9.93]
A8	24.31	17.34	-13.19 [-15.09, -11.30]	-10.76 [-12.58, -8.94]

Table 6: F1 and EM on MuSiQue ($n=2,417$ questions paired) with paired-bootstrap 95% CIs on the delta vs NL ($n_{\text{boot}} = 10,000$, seed 0). EM and F1 deltas are direction-consistent across all five comparators. Note that TE's EM advantage over NL is small (+1.08 pp) and its 95% CI crosses zero, whereas its F1 advantage (+5.24 pp) is clearly positive. The three trivial-compression controls (A6/A7/A8) show large negative deltas on both metrics with CIs well below zero, and LLMLingua-2 is negative on both metrics with CIs excluding zero.

The direction-consistency across F1 and EM lends confidence that the C2 mechanism finding (A6/A7/A8 all far below TE at matched budget) is not an F1-scorer artefact. On the positive TE >NL comparison, the EM CI that crosses zero is informative: it suggests TE's aggregate advantage on MuSiQue is at least partly carried by partial-credit rewards where TE emits the correct bridge entity alongside additional tokens that F1 counts as recall but EM counts as a miss. This is consistent with the compression-where-redundancy interpretation in §6.1 (which is about mechanism at matched budget, not about the size of the aggregate F1 advantage). A full LLM-judge sensitivity sweep remains future work (§8.1).

H Pre-registration protocol and amendments

These pre-commitments—FDR-BH correction across {MuSiQue, 2Wiki, meta}, a heterogeneity rule ($I^2 > 75\%$; it did not fire), an analytic MDES bound reported alongside the null, and explicit hypothesis-generating-only labelling of the between-dataset moderator—jointly prevent auxiliary-to-primary promotion after the null primary.

The pre-registration was filed prior to data collection (DOI and filing date omitted from the submission version for reviewer anonymization; restored in the camera-ready). The primary C1' hypothesis, the auxiliary C2 mechanism observation (§11.7 obs 2), the FDR-BH correction family, the DerSimonian–Laird pooling specification, the pre-committed $I^2 > 75\%$ heterogeneity rule, and the landmark kill-gate tests K-F1-1, K-TA-1, and K- γ -1 are all pre-registered. Two in-repo amendments were filed during the study:

- §10 amendment: documented a default-prompt bug discovered in the Wave-1a pilot and specified the remedial Wave-1b rerun at matched LLMingua-2 budget.
- §11 amendment: filed C2 (compression-where-redundancy, §11.7 obs 2) as a separately pre-registered auxiliary mechanism, prior to the Wave-2 cloud run that produced the A6/A7/A8 data reported in Table 1.

A full amendment log with SHA-stamped timestamps is included in the supplementary materials.

I Post-hoc observation: between-dataset gradient

This appendix expands the brief main-body pointer at the end of §5.4 into the full detail of the between-dataset TE–NL gradient. We report the gradient as an *exploratory* observation only, not as a contribution.

The between-dataset ordering in Table 2 (MuSiQue +4.75 significant, 2Wiki +1.53 null, HotpotQA –2.34 significant negative) is direction-consistent with the pre-registered P2b verbatim:

“On HotpotQA (predominantly 1–2 hop), TE does *not* beat NL: $\Delta F1(NL - TE) > 0$. This validates that the interaction is emergent with depth, not a global TE-wins effect.” (pre-registered document, §4)

The direction is pre-registered; the specific moderator that *explains* the gradient is not. In our $n = 3$ datasets the ordering correlates monotonically with dataset NL F1 (MuSiQue 38.0, 2Wiki 56.0, HotpotQA 69.7). Hop-based moderators (modal hop count, mean hop depth, 4-hop share) do *not* have the monotone shape required: modal hop counts are 2/no-mode/2 respectively, mean hop depths are 2.65/3.0/2.0, and 4-hop shares are 17%/33%/0%. NL-ceiling-proximity is therefore the only monotone candidate moderator at $n = 3$. We report it as a post-hoc observational pattern only—unobserved dataset-construction factors (e.g., bridge-entity retrievability, distractor-passage content) correlated with NL ceiling cannot be ruled out. Figure 3 shows the three points with no fitted line and an explicit $n = 3$ caveat.

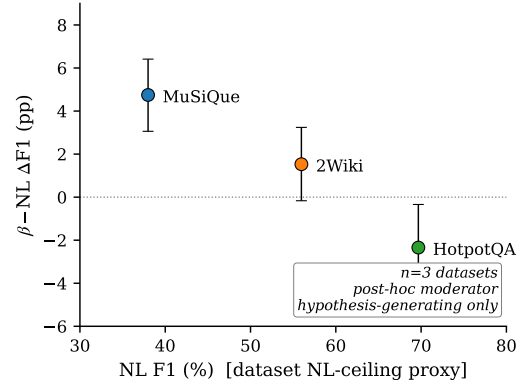


Figure 3: Post-hoc NL-ceiling-proximity observation ($n = 3$ datasets, hypothesis-generating only). Between-dataset TE–NL $\Delta F1$ (pp) plotted against dataset NL F1 (%). With only three datasets any monotonic moderator fits similarly; we therefore show no fitted line, no R^2 , and no coefficient. The ordering is consistent with—but does not confirm—an NL-ceiling-proximity reading; unobserved dataset-construction factors correlated with NL ceiling cannot be ruled out at $n = 3$. Error bars: 95% paired-bootstrap CI on $\Delta F1$.

At $n = 3$ datasets, NL-ceiling-proximity is the only monotone candidate moderator; unobserved dataset-construction factors (e.g., bridge-entity retrievability, distractor-passage content) cannot be ruled out. We therefore do not treat the gradient as a finding of this paper. A confirmatory replication with a fourth dataset where NL ceiling varies independently of dataset construction is the minimum prerequisite for any claim; we reserve that for follow-up work.

Hop distribution per dataset (reported for completeness, not a candidate moderator at $n =$

3). MuSiQue: 2-hop 1,252 (51.8%), 3-hop 760 (31.4%), 4-hop 405 (16.8%); modal 2, mean 2.65. 2Wiki: 2-hop 500, 3-hop 500, 4-hop 500 (balanced); no single mode, mean 3.0. HotpotQA: 2-hop 1,000 (100%); modal 2, mean 2.0. None of modal hop, mean hop, or 4-hop share is monotone with the between-dataset TE – NL gradient.

J F3 coherent-prose comparator: entity coverage and landmark contrasts

F3 evaluates Claude Sonnet 4.6 as a frontier coherent-summary encoder on the same three datasets, post-truncated to each per-row TE qwen-token budget (A7-style), so F3 is evaluated at the same budget as TE on identical consumer rows. Entity-coverage QC on a 10-row random MuSiQue sample (seed 0): at matched TE qwen-token budget F3 retains 0.497 of TE’s named entities, versus 0.782 for F3 raw (pre-truncation) at $1.59\times$ TE’s budget and 0.824 for LLMingua-2 rate-50 at $2.29\times$ TE’s character budget (Table 7); the drop $0.782 \rightarrow 0.497$ is produced entirely by matched-budget truncation, not by the encoder. Comparing F3 vs A7 on MuSiQue (both truncated to TE’s per-row qwen-token budget, differing only in summarization quality) gives F3–A7 $\approx +7$ pp: coherent summarization adds value over raw truncation even after both lose entity coverage, so the TE–F3 contrast in Table 3 isolates the density advantage from the summarization-quality advantage. Paired-bootstrap 95% CIs: $n_{\text{boot}} = 10,000$, seed 0, paired by `question_id`; per-dataset n : MuSiQue 2,417; 2Wiki 1,500; HotpotQA 1,000.

Compressor	Entity coverage (median)	Budget regime
TE (Telegraph English)	1.000 [†]	matched TE qwen-tokens
F3 (coherent prose, truncated)	0.497	matched TE qwen-tokens
F3 raw (pre-truncation)	0.782	$1.59\times$ TE qwen-tokens
LLMLingua-2 rate-50	0.824	$2.29\times$ TE char budget

Table 7: Entity-coverage QC on a 10-row random MuSiQue sample (seed 0). [†] TE is the reference by construction. F3 at matched qwen-token budget retains about half of TE’s named entities; the ~ 30 -pp gap from F3 raw is produced by matched-budget truncation, not by the Sonnet encoder. LLMingua-2’s higher coverage is measured at a different (more permissive) budget regime.

K Honest negatives: F2-struct and γ -rescue

Two pre-registered kill gates fired. F2-struct (a structured NL reformulation) does not improve

over NL on any dataset ($\Delta = -6.6, -13.4, -15.7$ pp on MuSiQue, 2Wiki, HotpotQA respectively); F2-struct is cut per pre-registration. K- γ -1, a hop-3 rescue frame intended to outperform TE on 3-hop questions in a revise-and-review pass, fails: combined MuSiQue + 2Wiki hop-3 $\Delta(\gamma - \text{TE}) = -0.15$ pp, 95% CI $[-0.78, +0.48]$, $n = 1,260$. γ is cut. Both negatives are useful: F2-struct falsifies a “surface structure alone suffices” hypothesis, and K- γ -1 falsifies a specific “deeper revision pass rescues the hard hops” hypothesis.

L Follow-up pilot: small-LM encoder on MuSiQue with oracle entity conditioning

Motivation. §6.3 chose Claude Sonnet 4.6 as the TE encoder to isolate the consumer-reading question from translator-capacity confounds. A reviewer may reasonably ask whether a much smaller encoder, given domain-matched entity-preserving training, could substitute. This pilot is a single-dataset *proof-of-concept* for encoder substitutability; it is not a full replication of the main-experiment matrix and does not close the general translator-capacity question.

Setup. We fine-tuned Qwen-3.5-0.8B (base) via LoRA on 5,000 entity-preserving Wikipedia QA rows drawn from the MuSiQue and HotpotQA training splits (2,500 each). Each training row pairs a question, a retrieval passage, the answer, and the decomposition’s *bridge entities* (the set of surface forms the decomposition labels as appearing in intermediate hops). The encoder is conditioned at both train and inference on the gold bridge-entity set as a system-prompt side channel; the supervised target is a Telegraph-English paraphrase of the passage produced by Sonnet 4.6 under a bridge-entity-aware prompt. At evaluation we re-translate the held-out MuSiQue $n=2,417$ test shard through the fine-tuned encoder under the same oracle conditioning, then send the output to the main Qwen-3.5-9B consumer using the same default prompt, decoder setting, and F1 scorer as the main experiments.

Results. Evaluating all four conditions on the same $n=2,417$ held-out items and pairing on question id:

- E10-v2 (retrained 0.8B): F1 48.04%
- TE (Sonnet 4.6, same-job): F1 42.40%

1042 • NL baseline: F1 37.50%
 1043 • LLMingua-2 (rate-50, matched to TE bud-
 1044 get): F1 35.07%

1045 Paired bootstrap over questions ($n_{\text{boot}} = 10,000$,
 1046 seed 0):

1047 • E10-v2 – LLMingua-2: +12.97 pp, 95%
 1048 CI [+11.00, +14.92]
 1049 • E10-v2 – TE: +5.64 pp, [+3.76, +7.54]
 1050 • E10-v2 – NL: +10.54 pp, [+8.55, +12.53]

1051 All three CIs are strictly positive. The same-
 1052 job TE F1 42.40% differs by 0.3 pp from the
 1053 main-experiment TE 42.74%, consistent with
 1054 deterministic-decode batch-order shifts across
 1055 consumer-eval runs.

1056 **Scope limits.** Three limitations bound the read-
 1057 ing of these numbers.

1058 (i) *Single-dataset.* The pilot tests only
 1059 MuSiQue; the main paper’s claims span three
 1060 datasets. Running the pilot on 2Wiki and Hot-
 1061 potQA is mechanical but was deprioritised as the
 1062 MuSiQue result already suffices to address the
 1063 “frontier-in, frontier-out” framing concern on the
 1064 dataset where TE most clearly dominates.

1065 (ii) *Budget mismatch.* The retrained 0.8B output
 1066 sits at $\sim 10\%$ of TE’s qwen-token budget (median
 1067 126 vs. 1251 per row; only 9/2417 rows exceed
 1068 TE’s budget and are truncated). The +5.64 pp
 1069 lift over TE therefore reads as an extreme-density
 1070 data point, not matched-budget dominance. A
 1071 matched-budget probe (letting E10-v2 generate
 1072 up to TE’s per-row budget) would disambiguate
 1073 *extreme-density is sufficient* from *E10-v2 happens*
 1074 *to work at low budget*; we do not run that probe
 1075 here.

1076 (iii) *Oracle conditioning at inference.* The de-
 1077 composition’s bridge-entity set is used as prompt-
 1078 ing input to the encoder at inference time. In
 1079 a deployment setting these labels are not avail-
 1080 able without either human annotation or a sepa-
 1081 rate entity-extraction model; the pilot thus demon-
 1082 strates an *upper bound* on encoder substitutability,
 1083 not a deployable alternative. A non-oracle control
 1084 (conditioning on NER-predicted bridge entities,
 1085 or on entities recovered from the consumer’s first-
 1086 pass guess) is the natural follow-up.

1087 **What the pilot does and does not establish.**
 1088 Subject to limits (i)–(iii), the pilot establishes
 1089 that on MuSiQue translator capacity is *not the*

binding constraint on our mechanism and den-1090
 sity claims: a domain-matched 0.8B encoder with 1091
 oracle entity conditioning reaches consumer F1 1092
 above LLMingua-2 and above the frontier TE 1093
 encoder at a fraction of the token budget. It does 1094
not establish that translator capacity is unimpor-1095
 tant across 2Wiki or HotpotQA, that small-LM 1096
 encoders match TE without oracle conditioning, 1097
 or that the density property transfers to matched-1098
 budget generation. The general translator-capacity 1099
 question remains open and is the subject of con-1100
 current work. 1101

Artifacts. The following pilot artifacts accom-1102
 pany this paper as supplementary material: the 1103
 5,000-row entity-preserving training data; the 1104
 Sonnet-generated Telegraph-English targets; the 1105
 LoRA fine-tune config and training-loss curves; 1106
 the re-translated MuSiQue $n=2,417$ test shard 1107
 produced by the fine-tuned encoder; the consumer-1108
 eval shard, per-row F1/EM outputs, and paired-1109
 bootstrap JSON with all three contrasts. 1110