AACL-IJCNLP 2026 · MAIN CONFERENCE · PAPER 330

Text-Preserving Lossy Text Compression

A Study of Strategic Deletion and LLM Reconstruction

Yuchun Zou[1] · **Junhong Tong**[2] · **Jun Li**[1 2]

[1] CUNY Graduate Center [2] CUNY Queens College

Code, data, and per-sample outputs are on GitHub:
github.com/junli-cuny/lossy-text-compression-benchmark

MOTIVATION

Many pipelines need the gist, not every byte back

- Lossless codecs such as zlib [1] keep every byte, so they save little on natural language. On BBC News, zlib still stores **51%** of the bytes.
- Many LLM pipelines must pass text through a narrow channel. A document may not fit the context window, and an agent's memory keeps growing. A weak device may also need to send text to a strong server.
- The encoder runs on **every** chunk, but the decoder runs only on the **few** chunks that are requested later.

What this setting needs

The encoder must be cheap.

The compressed form must stay plain text, so it can still be searched, edited, and zipped.

The reconstruction must stay faithful to the original.

No systematic benchmark existed for this **text-preserving lossy** setting.

THE TASK

The encoder deletes words, and an LLM rebuilds the text

1 Source text

The encoder deletes words, spaces, and punctuation by how common they are.

~~Film~~ stars ~~from~~ across the ~~globe are preparing~~ to walk the ~~red~~ carpet ~~at~~ this year's Bafta ~~award ceremony~~...The ~~2005 Orange~~ British ~~Academy Film~~ Awards ~~are~~ being ~~held~~ at The ~~Odeon in~~ London's Leicester Square~~. A host~~ of Hollywood stars~~,~~ including Cate Blanchett, Leonardo DiCaprio, ~~Keanu~~ Reeves and Richard ~~Gere~~, ~~are expected to~~ attend ~~Saturday's~~ ceremony~~.~~

The encoder deletes the struck-out text.

2 Skeleton

This is what gets stored. About half of the article remains, and it is still plain text.

```
starsacrossthe to walk the  carpet this year'sBafta...The  BritishAwards
beingatThe  Londons Leicester Square  of Hollywoodstars
includingCateBlanchett,Leonardo DiCaprio, Reeves andRichard ,    attendsceremony
```

The LLM decoder rebuilds the text.

3 LLM reconstruction

An LLM sees only the skeleton. Green words were deleted and came back exactly.

Stars **from** across the **globe are** set to walk the **red** carpet **at** this year's BAFTAs. The **EE** British **Academy** Awards **are** being **held** at The **Odeon in** London's Leicester Square, with **a host** of Hollywood stars, including Cate Blanchett, Leonardo DiCaprio, **Keanu** Reeves, and Richard **Gere**, **expected to** attend the ceremony.

Grammar brings back small words, and world knowledge brings back names such as Keanu Reeves and the Odeon. One slip remains, because “2005 Orange” became **“EE”**, a name the article never uses.

Why an LLM rebuilds the text

Deletion cannot be undone by a rule.

Many different texts shrink to the same skeleton. The decoder must guess the most likely original, and predicting missing text is what a language model is trained to do.

The deleted words are predictable.

WordFreq removes common words first, and Entropy removes low-surprisal tokens first. An LLM can usually predict these words from the ones that remain.

Pretraining adds knowledge of the world.

Names, events, and familiar phrasing help fill gaps that the skeleton leaves open. The same prior also produces fluent errors such as “EE” in the last example, so fidelity must be measured.

The decoder's cost is paid rarely.

It runs only on the few chunks that are read back, often on a strong server. One model serves every deletion strategy zero-shot, and a fine-tuned 3B model can run locally.

A model that predicts text well can also compress it [2]. Here the prediction happens at the receiver.

POSITIONING

Our target is fidelity to the original text

Summarization

It rewrites the source into a shorter text [3–5].

It keeps no skeleton, so there is nothing to reconstruct.

The original wording and names can be lost.

Prompt compression

Methods such as LLMLingua [6] and Selective Context [7] belong here.

They optimize one downstream task for a fixed consumer model.

This work

We keep original tokens as anchors.

We score the reconstruction by its **fidelity to the original**.

The skeleton is a reusable text artifact, not a one-off prompt.

Deletion strategies at three levels

Level 1 · characters

Needs no model.

Step deletion removes every k-th character. Its random variants pick positions by Gaussian, Bernoulli, or Poisson sampling.

Level 2 · words

Needs a lookup table.

WordLen edits words by shape in stages. It drops vowels and short words, then trims words to stems.

WordFreq-3 sorts words into three Zipf frequency classes [8] and splits the budget across them in proportion. It then drops whole words at random within each class. **WordFreq-6** does the same with six buckets, adding punctuation, other tokens, and whitespace.

LP uses six frequency buckets. A linear program sets how much to delete from each bucket to maximize estimated BERTScore.

Level 3 · surprisal

Runs GPT-2 [9] at the encoder.

Entropy drops the most predictable tokens first. **Entropy-LP** and **Hybrid-α** combine surprisal with the LP allocation or with frequency rank.

Comparisons

LLMLingua-2 [10], a trained token classifier from prompt compression, is tested as a drop-in encoder. **Summarization** to the same length is a baseline with no skeleton.

SETUP

One benchmark, several lenses

Decoders

Google Gemini Flash [11, 12] decodes zero-shot.

We fine-tune Llama-3.2-3B [13] with QLoRA [14], and Qwen2.5-3B [15], so decoding can run locally.

Retention

We test 9 rates, from 0.9 down to 0.1.

A rate of 0.05 serves as a stress test.

Metrics

We score with BERTScore [16], ROUGE-L [17], and entity recall.

People judge facts blind, and two LLM raters [18, 19] scale up their judgments.

Data

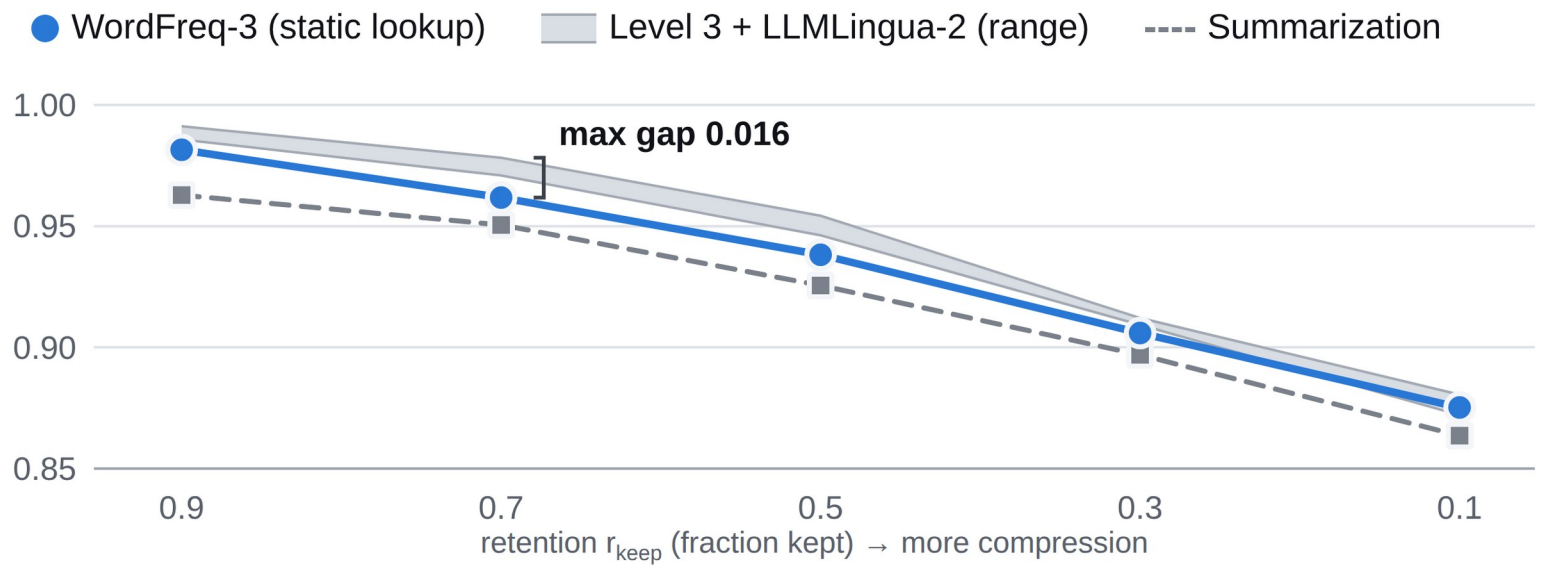
BBC News [20] is the main corpus.

WikiText [21] and Reddit [22] add other English domains.

Three Chinese corpora cover news [23], Wikipedia [24], and Zhihu [25].

FINDING 1

WordFreq is a strong, nearly free encoder



~1 ms
WordFreq lookup

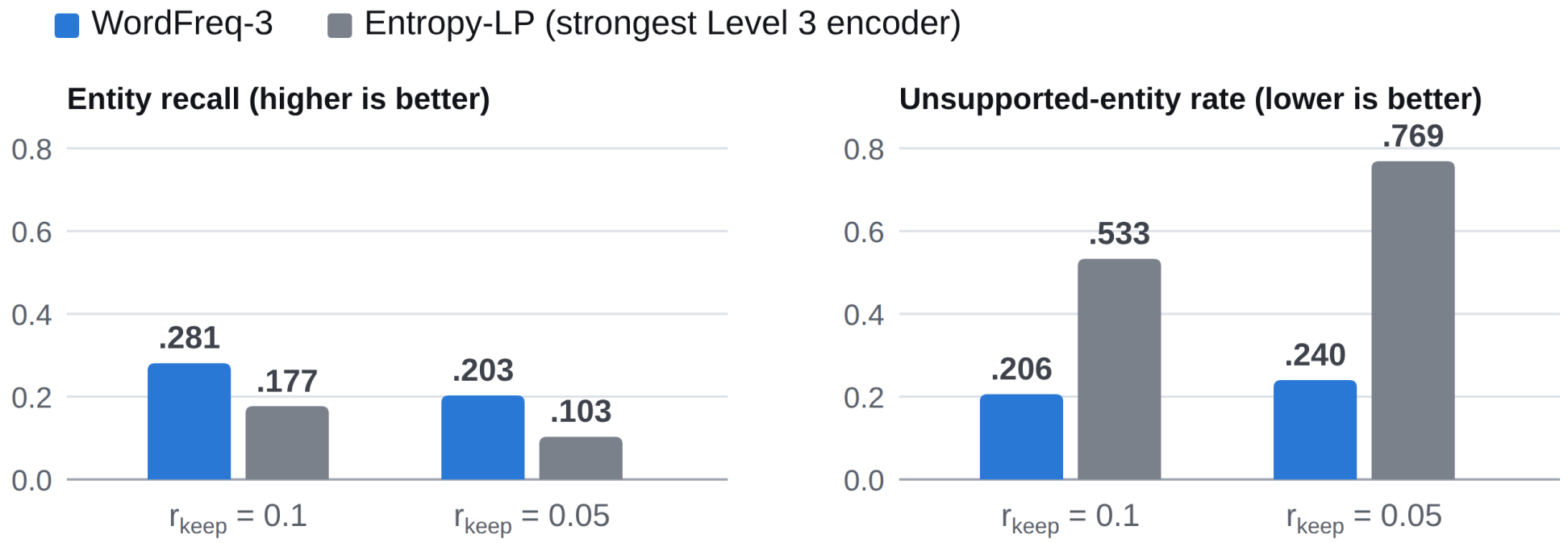
~15 ms
GPT-2 surprisal, GPU

~130 ms
GPT-2 surprisal, CPU

The gray band spans the three **Level 3 encoders** (Entropy, Entropy-LP, and Hybrid-0.5) and LLMLingua-2. The gap to WordFreq is largest at mild-to-moderate compression. By r_{keep} = 0.3 it is about 0.006, and at 0.1 WordFreq falls inside the band.

FINDING 2

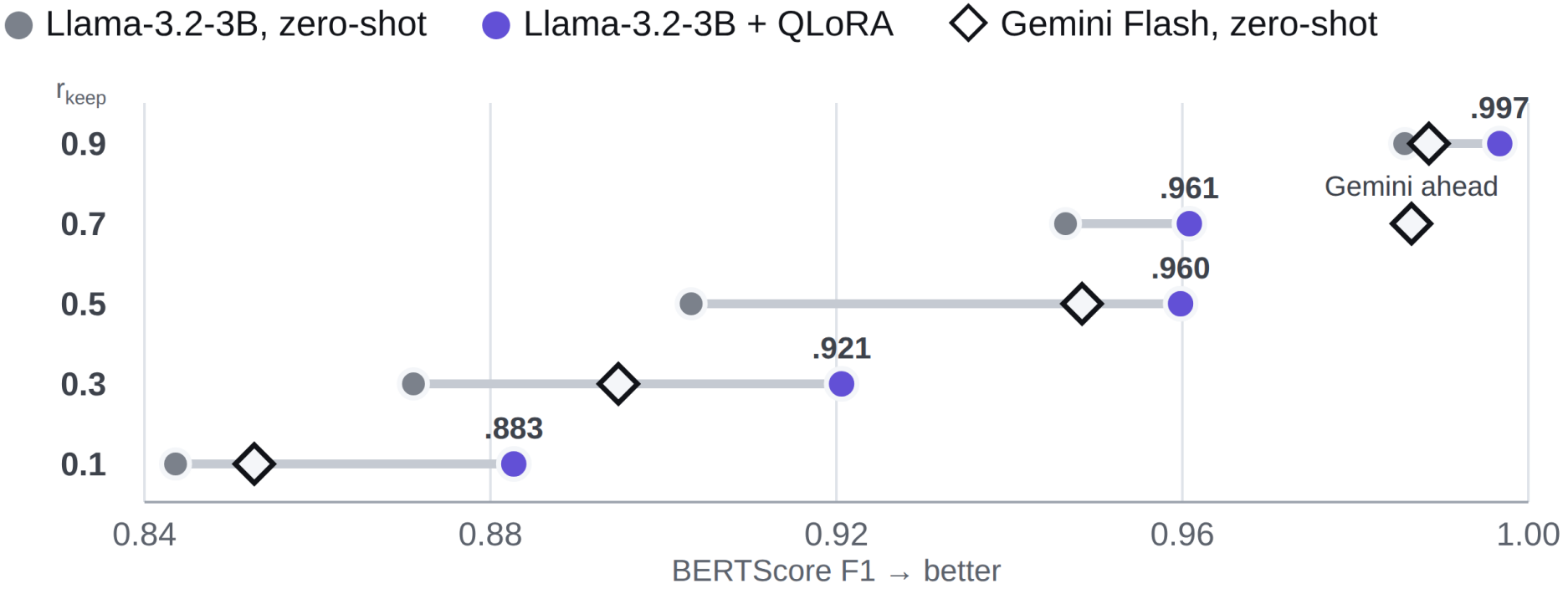
At low retention, WordFreq keeps more entities



At r_{keep} = 0.05, WordFreq keeps about **twice** as many of the original entities as Entropy-LP, at about **a third** of its unsupported-entity rate.

FINDING 3

A fine-tuned 3B decoder rivals Gemini



The QLoRA model beats zero-shot Gemini Flash at **4 of 5** rates. With the skeleton fixed, fine-tuning adds 0.011 to 0.056 BERTScore, so reconstruction can run fully on a local machine.

EVALUATION

BERTScore overstates what survives

In a blind human study at r_{keep} = 0.1, people judged WordFreq-6 and LP reconstructions.

≈0.84

BERTScore F1

It looks close to the source.

0.14

ROUGE-L F1

Most of the wording is gone.

3%

of facts preserved

This is what people judged.

- Similarity still tracks facts overall (Spearman ρ = 0.776 over 54 items), but stays high at low retention, where fluent text replaces lost content.
- Only 4 of 54 items invented a claim outright. Most losses are omissions and distortions.

When you evaluate reconstruction, also report word overlap (such as ROUGE-L) and fact preservation (such as entity recall).

IN PRACTICE

Choosing an operating point

Regime	r_{keep}	Stored*	Entity recall	Unsupported	Suitable for
Mild	0.9–0.8	0.43–0.50	0.89–0.95	0.03–0.06	archival, agent memory
Moderate	0.7–0.5	0.29–0.39	0.65–0.88	0.07–0.15	gist: logs, previews
Aggressive	0.3–0.1	0.08–0.21	0.19–0.54	0.16–0.26	a stress test, not an operating point

*Stored is the skeleton after zlib, as a fraction of the original bytes. zlib alone stores 0.513 of the original. Each band spans WordFreq and LP on BBC News.

Rule of thumb

Start with WordFreq. It needs no encoder model, and it is the most robust choice once $r_{keep} \leq 0.3$.

Switch near $r_{keep} \approx 0.5$. There, LP or a surprisal hybrid can pay off if a model can run at the encoder.

CONCLUSION

Takeaways

Beyond BBC News

We repeated the study on WikiText, Reddit, and Chinese news, Wikipedia, and Zhihu.

Level 1 (Step deletion) is weak in every domain once compression is substantial. WordFreq stays a strong low-cost baseline.

The best rule shifts with the domain. Re-check on your own data, and report word overlap and fact preservation too.

1. WordFreq, a static word-frequency lookup, is a strong, nearly free encoder.

2. Level 3 encoders help most at mild-to-moderate compression. At the lowest retention, WordFreq keeps more entities.

3. A QLoRA-tuned 3B decoder rivals zero-shot Gemini Flash.

4. BERTScore alone overstates fidelity, so also report word overlap (ROUGE-L) and fact preservation (entity recall).

Thank you! Code, data, and per-sample outputs are on GitHub:
github.com/junli-cuny/lossy-text-compression-benchmark

yzou@gc.cuny.edu · jun.li@qc.cuny.edu

References

[1] Gailly, J. and Adler, M. 1995. zlib compression library. zlib.net.

[2] Delétang, G. et al. 2024. Language modeling is compression. ICLR.

[3] Rush, A. M., Chopra, S., and Weston, J. 2015. A neural attention model for abstractive sentence summarization. EMNLP.

[4] See, A., Liu, P. J., and Manning, C. D. 2017. Get to the point: Summarization with pointer-generator networks. ACL.

[5] Liu, Y. and Lapata, M. 2019. Text summarization with pretrained encoders. EMNLP-IJCNLP.

[6] Jiang, H. et al. 2023. LLMLingua: Compressing prompts for accelerated inference of large language models. EMNLP.

[7] Li, Y., Dong, B., Guerin, F., and Lin, C. 2023. Compressing context to enhance inference efficiency of large language models. EMNLP.

[8] Speer, R. 2022. wordfreq: Access a database of word frequencies in multiple languages. Version 3.0.

[9] Radford, A. et al. 2019. Language models are unsupervised multitask learners. OpenAI technical report.

[10] Pan, Z. et al. 2024. LLMLingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. Findings of ACL.

[11] Google DeepMind. 2025. Gemini 2.0 Flash model card.

[12] Gemini Team, Google. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv:2507.06261.

[13] Llama Team, AI @ Meta. 2024. The Llama 3 herd of models. arXiv:2407.21783.

[14] Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. 2023. QLoRA: Efficient finetuning of quantized LLMs. NeurIPS.

[15] Qwen Team. 2024. Qwen2.5 technical report. arXiv:2412.15115.

[16] Zhang, T. et al. 2020. BERTScore: Evaluating text generation with BERT. ICLR.

[17] Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. Text Summarization Branches Out.

[18] Anthropic. 2026. System card: Claude Opus 5.

[19] OpenAI. 2026. GPT-5.6 system card.

[20] Kalpande, G. 2022. BBC news summary dataset. Hugging Face.

[21] Merity, S., Xiong, C., Bradbury, J., and Socher, R. 2017. Pointer sentinel mixture models (WikiText). ICLR.

[22] MentionBroker. 2026. Reddit comment generation dataset. Hugging Face.

[23] Ge, Q. 2018. Chinese official daily news. Kaggle.

[24] BigScience Data. 2022. ROOTS Chinese (traditional) Wikipedia dataset. Hugging Face.

[25] Wang, R. 2023. Zhihu-KOL. Hugging Face.