Concept than Document: Context Compression via AMR-based Conceptual Entropy

Anonymous ACL submission

Abstract

Large Language Models (LLMs) face information overload when handling long contexts, particularly in Retrieval-Augmented Generation (RAG) where extensive supporting documents introduce redundant content that interferes with reasoning. Context engineering has emerged to address these challenges, yet existing methods rely on lexical or token-level features that fragment semantic units and fail to capture conceptually essential content. We propose an unsupervised context compression framework leveraging Abstract Meaning Representation (AMR) to preserve semantically essential information while filtering irrelevant text. By quantifying node-level entropy within AMR graphs, our method estimates the conceptual importance of each node, enabling retention of core semantics. Specifically, we construct AMR graphs from retrieved contexts, compute the conceptual entropy of each node, and identify statistically significant concepts to form a condensed, semantically focused context. Experiments on the PopQA and EntityQuestions datasets demonstrate that our method outperforms vanilla RAG and existing baselines, achieving superior accuracy while substantially reducing context length. To the best of our knowledge, this is the first work introducing AMR-based conceptual entropy for context compression, demonstrating the potential of structured linguistic representations in context engineering.

1 Introduction

Large Language Models (LLMs) are increasingly equipped with mechanisms to incorporate long contexts, allowing them to leverage external information beyond their training data (Lewis et al., 2020; Karpukhin et al., 2020). However, as the context length grows, LLMs often struggle to effectively identify and utilize truly relevant information, leading to performance degradation and inefficiency. This challenge, reflecting the trade-off between retrieval recall and precision, becomes particularly acute in scenarios such as Retrieval-Augmented Generation (RAG), where the inclusion of more retrieved documents raises the chance of accessing useful knowledge but simultaneously introduces overwhelming amounts of irrelevant text obscure the key facts (Shi et al., 2023; Jin et al., 2025a).

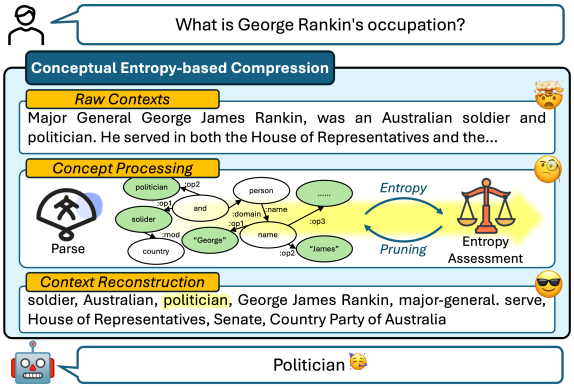

Figure 1: Long retrieved documents contain much irrelevant content; our method keeps only key AMR-based concepts to form a semantically focused context.

Context engineering has therefore become an effective strategy for enhancing the quality and efficiency of long-context utilization, aiming to distill essential information while reducing noise and redundancy (Mei et al., 2025). Existing approaches primarily focus on lexical or surface-level features for information filtering (Xu et al., 2024; Cheng et al., 2024). While these methods work well for certain queries, they may struggle with capturing complex semantic concepts and preserving factually important information. Moreover, traditional compression techniques may inadvertently remove crucial supporting evidence while retaining superficially relevant but semantically vacuous content.

To address the aforementioned limitations, we propose a novel context compression method that leverages Abstract Meaning Representation (AMR) (Banarescu et al., 2013) to identify and preserve semantically essential information. AMR

graphs provide a structured representation that abstracts away from surface syntactic variations while retaining core semantic content (Chen et al., 2025). Concepts assuming diverse semantic roles across contexts naturally carry more informative value for inference (Kuhn et al., 2023), which can be quantified as higher information entropy in the role distribution of concept nodes (Nguyen et al., 2025). Moreover, cognitive studies suggest that the human brain can automatically reconstruct scenarios implied by essential concepts through pre-learned semantic knowledge (Binder et al., 2009; Horikawa, 2025), and LLMs exhibit a similar capacity for concept-based scene understanding, providing theoretical support for prioritizing semantically fundamental concepts during reasoning (Du et al., 2025).

Building on this foundation, our method constructs AMR graphs from retrieved contexts to represent entities and key semantic concepts in a structured form. For each concept node, we calculate the information entropy to estimate its semantic contribution based on its inherent contextual uncertainty. We then apply significance testing to identify concept nodes that exhibit statistically reliable informational salience, which constitute the structural basis for reconstructing a compressed context that preserves essential semantic content while suppressing redundant information. To mitigate potential distortion caused by AMR's abstraction from surface realization, the distilled concepts are restored to their original textual expressions in the source contexts, ensuring factural consistency and maintaining semantic clarity in the reconstructed compressed context for reasoning.

We evaluate our method on two challenging knowledge-intensive Q&A benchmarks, PopQA (Mallen et al., 2023) and EntityQuestions (Sciavolino et al., 2021), which require reasoning over long-context factual evidence retrieved from external sources. Experimental results show substantial performance gains over vanilla RAG and other context compression baselines, with more pronounced improvements on instances involving long supporting documents. These findings support our hypothesis that AMR-based entropy filtering effectively isolates core semantic content while removing redundant information. The main contributions of this work can be summarized as follows:

- We propose a novel unsupervised context compression framework that leverages AMR to identify and preserve core semantic information while filtering redundant content.
- Extensive experiments demonstrate that the proposed method outperforms vanilla and other compression baselines by maintaining robust semantic core preservation.
- The method achieves reductions in context length and latency while preserving semantic integrity, offering a linguistically empowered framework for context engineering.

2 Related Work

2.1 Context Engineering

Context engineering has become a key strategy for managing and structuring information in LLM workflows (Mei et al., 2025; Verma, 2024; Shi et al., 2024). Early approaches selected relevant sentences or passages based on lexical similarity (Hwang et al., 2024), while other methods used neural models to reorganize retrieved contexts (Xu et al., 2024; Liu et al., 2024). Recent work examines learned context engineering techniques that optimize representations for downstream tasks. Jiang et al. (2024) uses instruction tuning to refine contexts while preserving task-relevant information. Selective-Context (Li et al., 2023b) applies attention mechanisms to highlight critical segments. Jin et al. (2025b) emphasizes semantic integrity in engineered contexts, integrating natural language spans and semantic vectors to support dynamic evidence selection and improve answer quality.

2.2 AMR-enhanced Large Language Models

Abstract Meaning Representation provides a structured formalism that abstracts away from syntactic variations, making it suitable for cross-lingual and cross-domain applications (Wein and Opitz, 2024). Recent AMR parsing advances have made it practical to construct high-quality graphs from context (Bevilacqua et al., 2021; Zhou et al., 2021), enabling applications across NLP tasks(Li et al., 2021; Liu et al., 2015; Song et al., 2019). With the rise of LLMs, researchers have explored using AMR for semantic enhancement. Recent studies have examined AMR-driven chain-of-thought prompting, showing that structured semantic representations can improve LLM performance across tasks (Jin et al., 2024). Zhang et al. (2025) integrates AMR into LLM frameworks through structured representation methods, although aligning

AMR's graph structure with sequential processing remains challenging. These observations suggest that while full graph utilization is non-trivial, AMR nodes constitute stable semantic units that encode informative conceptual content, making them suitable for structured context engineering.

2.3 Information Theory in LLMs

Information-theoretic measures have become increasingly important in the era of LLMs, providing principled tools to understand and improve model behavior (Wang et al., 2025). LLMs have leveraged such analyses for interpretation and optimization (Nikitin et al., 2024). For instance, entropy-based selection of demonstration examples has been shown to enhance the performance of CoT prompting (Zhou et al., 2023). Beyond prompting, information-theoretic approaches have been applied to model compression, knowledge distillation, and efficient fine-tuning (Yin et al., 2024; Mao et al., 2024). These studies illustrate an emerging trend in which information theory provides both theoretical insights and practical tools for working with LLMs (Agarwal et al., 2025). In this work, we integrate graphical information-theoretic principles of AMR, leveraging high-entropy nodes as concise and informative representations of long contexts.

3 Methodology

3.1 Problem Formulation

The framework for transferring the context in raw documents to condensed concepts is as Figure 2. Given a query Q and a set of retrieved documents $D = \{d_1, d_2, ..., d_n\}$ with corresponding correct answers $A = \{a_1, a_2, ..., a_m\}$, our objective is to generate a compressed context C' that preserves the most semantically informative concepts essential for answering Q to yield $a_j \in A$, while substantially reducing the overall context length.

To create a controlled experiment that focuses exclusively on the impact of core concepts within the context on answer accuracy, we retain only documents that contain correct answers. This controlled setting enables us to isolate how our compression method affects the preservation of essential contextual information by eliminating interference from irrelevant documents. The hypothesis can be formalized as: $\forall d_i \in D, \exists a_j \in A$ such that $a_j \in d_i$.

Formally, we aim to learn a compression function $f(D) \rightarrow C'$ such that:

$$Acc(q, C') \gtrsim Acc(q, D) \text{ and } |C'| \ll |D| \quad (1)$$

where $C' \subseteq D$, $|C'|$ and $|D|$ are the lengths of the compressed and original contexts, respectively.

3.2 AMR Graph Construction

For each document $d_i \in D$, we construct it to the sentence-level AMR graphs with an mBart-based parser [1] trained in the AMR 3.0 corpus [2] to address potential multilingual concerns. Let $G_i = (V_i, E_i)$ denote the AMR graph for document d_i, where V_i represents the set of concept nodes and E_i represents the semantic relations between concepts. Each concept node $v \in V_i$ corresponds to a semantic concept (e.g., entities, predicates, or modifiers) and is associated with its textual realization in the raw document. The edges in E_i represent semantic relationships such as agent-of (ARG0), patient-of (ARG1), and various semantic roles.

Our approach is grounded in the cognitive hypothesis that both human comprehension and LLM inference can effectively reconstruct semantic scenarios from discrete informative concepts without explicit relational encoding (Xu et al., 2025; Fedorenko et al., 2024; Rogers et al., 2004; Wit and Gillette, 1999). This principle suggests that intelligent systems possess inherent capabilities to infer implicit relationships between concepts based on their learned background knowledge and contextual co-occurrence patterns (Brown et al., 2020; Cao et al., 2023; Suresh et al., 2023). Building on these foundations, we keep the concept nodes V_i and discard the explicit E_i in each G_i. This design ensures that the compressed context consists of discrete semantic concepts, avoiding the introduction of artificial relational symbols that may interfere with the LLM's pre-trained language understanding capabilities while leveraging the model's intrinsic ability in concept-based scenario reconstruction.

3.3 Information Entropy Computation

To identify the most informative concepts within each AMR graph, we employ an information-theoretic approach based on token-level perplexity measurements. For each concept node $v \in V_i$, we calculate its information entropy by leveraging the AMR generation model's uncertainty when predicting the concept token sequence.

Given the AMR parsing model M with parameters θ, we obtain the probability distribution over the vocabulary for each token position in the

[1] https://github.com/BramVanroy/multilingual-text-to-amr

[2] https://catalog.ldc.upenn.edu/LDC2020T02

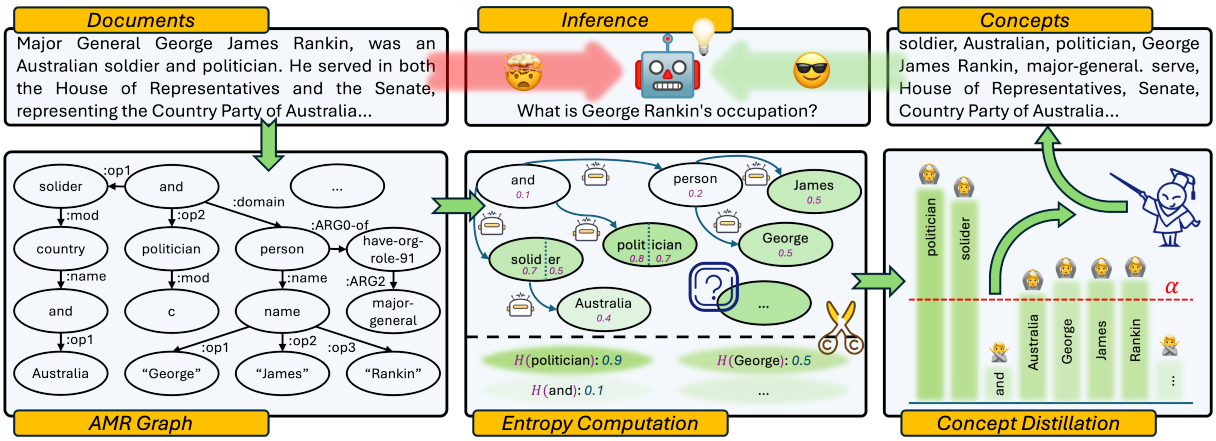

Figure 2: The conceptual entropy-based workflow converts the sparse context in raw supporting documents into condensed AMR-based concepts, forming a compact semantic representation for LLMs inference.

AMR linearization. However, modern tokenizers decompose words into subword units, requiring the aggregation to obtain concept-level entropy scores. For a concept v that corresponds to a complete word-level representation in d_i, the tokenizer may decompose it into the subword tokens $v = [s_1, s_2, ..., s_m], m \geqslant 0$. We compute the token-level entropy for each subword as:

$$E(s_j) = \exp\left(-\log P_\theta(s_j | s_{<j}, G_i)\right) \quad (2)$$

where $s_{<j}$ denotes the preceding tokens within the same concept. We aggregate token-level entropies into a concept-level entropy score as Eq. 3. Specifically, we identify concept boundaries by tracking tokens that begin with the special prefix "Ġ" and accumulate entropy scores for all s_j belonging to the same conceptual unit. This aggregation strategy ensures that concepts composed of multiple subword tokens are not artificially penalized relative to single-token concepts. This alignment provides a balanced representation of the model's uncertainty across all subword components of a concept.

$$H(v) = \frac{1}{m} \sum_{j=1}^{m} E(s_j) \quad (3)$$

Compared to token-level entropy in linear text, computing entropy over AMR concept nodes leverages semantic structure to more precisely estimate informational content. High-entropy nodes often represent content-specific, less redundant meanings, thus providing more discriminative signals for downstream reasoning. This enables the compression process to highlight semantically rich units that may be obscured in the surface text.

3.4 Concept Distillation

The supporting document set D can be conceptualized as a coherent descriptive scenario corresponding to query Q, within which genuinely informative concepts can be identified through their statistically significant entropy deviations. Concepts exhibiting higher entropy relative to the general nodes carry more discriminative information and are thus more valuable for answering the query. For each $d_i \in D$ with concept entropy $\{H(v_1), H(v_2), ..., H(v_{|V_i|})\}$, we perform a one-sample t-test to identify concepts with significantly higher information than the population mean:

$$t_{stat}(v_j) = \frac{H(v_j) - \bar{H}}{\frac{s}{\sqrt{n}}} \quad (4)$$

where $\bar{H}$ is the sample mean entropy, s is the sample standard deviation, and $n = |V_i|$. We compute the corresponding p-value using the t-distribution with $n - 1$ degrees of freedom:

$$p(v_j) = 2 \times (1 - F_t(|t_{stat}(v_j)|, n - 1)) \quad (5)$$

where F_t is the cumulative distribution function. We then screen out concepts whose p-values satisfy $p(v_j) < \alpha$ as statistically significant high-information concepts. Our goal is not to identify only the most informative concepts, but rather to eliminate overly generic ones while preserving a relative conceptual basis for LLMs to infer the document's semantics. Considering the empirical validation of LLMs' inference, we adopt a relaxed threshold, $\alpha = 0.3$. This setting prevents the over-pruning of moderately informative concepts,

thereby ensuring that the retained set includes contextual signals. The ablation study to verify the different α settings is in Section C.

3.5 Context Compression and Reconstruction

The final compressed context C' is constructed by aggregating the concepts with significant entropy across all documents in D. For each document d_i, let $V_i = \{v \in V_i : p(v) < \alpha\}$ denote the set of statistically significant concepts. For each $c_i \in C'$, the compressed representation for document d_i is:

$$c_i = \bigodot_{v \in V_i} \phi(v) \tag{6}$$

where $\phi(\cdot)$ maps each concept v to its processed surface form through a sequence of linguistic post-processing steps designed to preserve semantic coherence and ensure linguistic fluency. These include *Temporal Expression Reconstruction*, where date and time expressions fragmented during AMR parsing are converted into natural language format, such as transforming "month 7 year 2025" into "July 2025"; *Redundancy Removal*, which eliminates consecutive duplicate concepts to reduce repetition while maintaining semantic diversity; and *Surface Realization*, which restores the processed concepts to their original textual forms in the raw document to mitigate potential distortions introduced by the AMR parsing process. This compressed form serves as the final input context, preserving the essential semantic signals while substantially reducing the original context length.

4 Experiments

4.1 Datasets and Implementation Details

We conduct comprehensive evaluations on two widely-adopted open-domain question-answering datasets that provide long-context supporting documents for RAG-based inference: **PopQA** (Mallen et al., 2023) and **EntityQuestions** (Sciavolino et al., 2021). For comprehensive evaluation, we use Contriever (Izacard et al., 2022) as the retriever for PopQA and BM25 (Robertson et al., 2009) for EntityQuestions, with retriever optimization beyond the scope of this work. Both datasets are equipped with ground-truth annotations indicating whether each supporting document contains the correct answer, denoted by the boolean indicator "*hasanswer*". To align the problem formulation in Eq. 1, we retain only documents where "*hasanswer*" = *True*, ensuring that performance variations stem from compression effectiveness rather than irrelevant document interference. For each query Q, let K denote the number of answer-containing documents in the filtered D. The statistical characteristics of the curated $\langle Q, A, D \rangle$ triplets are summarized as follows:

Table 1: Statistical results of the amount of screened-out $\langle Q, A, D \rangle$ pairs from the datasets.

K=	1	2	3	4	5	6	7	8	9	10
PopQA	280	298	174	172	160	153	149	155	135	125
EQ	489	572	373	295	239	199	179	169	130	113

To mitigate reliance on parametric knowledge in LLM inference, we employ a structured prompting that prioritizes externally provided evidence over internal memory. We adopt the instruction as follows: "[*Refer to the following facts to answer the question. Facts:* C'. *Question:* Q]". Given that prompt intensity significantly influences inference behavior (Wu et al., 2024), we frame the supporting concepts C' as "facts" to establish a constrained knowledge boundary that minimizes interference from potentially conflicting parametric knowledge.

4.2 Baseline Methods

Our baseline evaluation examines two key dimensions: (1) diverse backbone LLM architectures, and (2) alternative context compression techniques. For backbone LLMs, we select mainstream publicly available LLMs, including GPT-Neo (1.3b and 2.7b) (Black et al., 2021), OPT (1.3b and 2.7b) (Zhang et al., 2022), BLOOM LM (560m, 7b1) (Le Scao et al., 2022), LLaMA-2-chat (13b) (Touvron et al., 2023), Llama-3.1-Instruct (8b) (Dubey et al., 2024), DeepSeek-V2-Lite (16b) (DeepSeek-AI, 2024), and Qwen3 (32b) (Team, 2025). The combination of backbone LLMs with contexts in raw supporting documents constitutes the *Vanilla* baseline.

For context compression, we implement five representative approaches that span different paradigms. We categorize these methods into three groups to answer the following questions: **Q1:** Can simple frequency-based measures suffice for identifying informative content? (*Statistical Method*). **Q2:** Can LLMs perform compression effectively through prompt-based reasoning? (*LLMs-driven Methods*). **Q3:** Can dedicated context compression models be more targeted and effective? (*Compression-specific Methods*).

The baselines corresponding to the above questions are as follows: (1) *Statistical Method*: TF-IDF, the statistical entropy-inspired method that

identifies salient terms using frequency–inverse document frequency weighting to highlight informative concepts. (2) *LLMs-driven Methods*: prompt-based keyword extraction and summarization that leverage LLaMA-3.1-8B-Instruct with prompts as Prompt A1 and Prompt A2 to generate keywords and summarizations. (3) *Compression-specific Methods*: Selective Context (SelCon) (Li et al., 2023a) that employs trained models to identify relevant spans, and LLMLingua (Jiang et al., 2023) uses budget-constrained token selection for optimal compression. These baselines evaluate if compressed contexts can preserve essential information while reducing computational overhead.

4.3 Evaluation Metrics

We employ two evaluation metrics: accuracy (Acc) and Area Under the Curve (AUC). The standard deviation (σ) of AUC is used as an auxiliary metric. The Acc follows the exact match protocol of Mallen et al. (2023), measuring if any generated answer exactly matches any gold-standard $a_j \in A$ for a given query Q. The σ assesses the stability of compressed methods across different backbone LLMs.

The AUC provides a comprehensive assessment across varying K. Specifically, AUC computes the area under the Acc curve against K. Higher AUC indicates superior overall performance across the corresponding intervals. Given our focus on long-context compression, we partition the AUC calculation into two intervals for the values of K: a standard interval $I_s = [1, 10]$ that captures general performance trends and a long-context interval $I_l = [6, 10]$ that highlights performance under long context. This decomposition provides clear insights into both typical and challenging scenarios.

5 Results and Analysis

5.1 Overall Performance

The AUC results in I_s interval in Table 2 and Table 3 present the overall performance comparison across both datasets. The full results in Acc are in Table A1 and A2 respectively. In the PopQA dataset, the proposed method achieves substantial gains compared to the vanilla baseline. The most notable improvements occur in larger models like Qwen3-32B, Llama-2-chat-13b, and DeepSeek-V2-Lite. In contrast, smaller models like Bloom-560m/7b1 show relatively modest improvements. On the EntityQuestions dataset, the results exhibit similar trends with some variations. The proposed method achieves the best or second-best performance across most configurations, with particularly strong results on larger models like Qwen3-32B. However, we observe slight performance degradation compared to vanilla on smaller models like GPT-Neo-1.3B and Bloom-560m/7b1. Considering the previous observation, this phenomenon indicates that compact LLMs may benefit from more contextual information that retains rich linguistic elements to reconstruct scenarios rather than aggressive compression. This suggests a trade-off between compression ratio and model capacity that warrants consideration in practical deployments. In addition, our method achieves a competitive σ across diverse backbone LLMs, indicating it preserves universally shared semantic cores rather than model-specific preferences, forming a robust semantic compression that maintains coherent reasoning chains across different architectures.

Compared to compression baselines, our method demonstrates substantial advantages across different paradigms. Against the statistical TF-IDF approach, we achieve overwhelming superiority on both datasets, outperforming all backbone LLMs. Although TF-IDF outperforms the vanilla setting on certain backbone models, this improvement is not consistent when examined across different architectures, as indicated by the unstable results with the highest σ. Its performance depends on surface-level lexical patterns, which may occasionally align with answer-bearing spans in simple contexts. However, TF-IDF lacks semantic structure awareness and does not model how LLMs reconstruct contextual meaning. As a result, it may either discard essential cues or retain redundant tokens that vary across models. The fluctuating performance across backbones indicates answers of **Q1** that frequency-based signals are insufficient for reliably identifying informative content.

The LLM-driven baselines, Keywords and Summary, show limited performance in most settings. Unlike statistical measures, these baselines depend on generative rewriting, which makes them sensitive to semantic integrity and prompts. These factors lead to unreliable results across different backbones. In addition, the generative paradigm can introduce hallucinations into the rewritten content, further increasing the uncertainty of the compressed context. A notable trend is that the summary-based compression achieves the lowest σ. The reason is summary-compressed context is remain natural language, forming a continuous repre-

Table 2: The quantitative results of AUC ↑ for the PopQA dataset, where the full name order of the LLMs is: GPT-Neo-1.3B, GPT-Neo-2.7B, OPT-1.3b, OPT-2.7b, Bloom-560m, Bloom-7b1, Llama-2-chat-13b, Llama-3.1-8B-Instruct, DeepSeek-V2-Lite, Qwen3-32B. The standard division is as $\sigma \downarrow$. The best results are in **bold**, and the second-best results are in underlined. The increased and decreased Δ are marked differently.

D	K	G-1.3	G-2.7	O-1.3	O-2.7	b-560	b-7b1	L-13	L3.1-8	DS-V2	Q3-32	$\sigma \downarrow$
Vanilla	I_s	553.32	550.79	585.12	596.31	575.04	664.92	583.57	701.36	575.00	251.99	119.63
	I_l	262.07	252.04	278.86	282.63	284.04	318.37	293.42	337.14	303.30	101.33	64.77
TF-IDF	I_s	354.04	508.48	486.22	523.84	417.67	608.85	623.00	650.98	179.28	210.62	165.39
	I_l	169.82	251.12	244.02	269.09	217.52	307.70	311.47	316.14	106.47	113.97	78.00
Keywords	I_s	423.52	449.40	532.66	547.01	497.93	588.64	552.55	606.34	295.62	271.88	116.40
	I_l	193.41	211.08	264.65	274.44	252.10	294.34	278.92	302.44	173.88	141.73	55.10
Summary	I_s	433.24	459.55	540.52	504.34	527.49	577.91	482.79	551.42	491.56	285.17	**82.74**
	I_l	206.04	223.84	267.55	242.91	268.18	294.93	252.27	270.41	269.50	138.74	44.81
SelCon	I_s	453.31	490.44	580.08	581.62	443.08	634.40	637.20	717.74	557.43	293.10	121.98
	I_l	209.18	228.22	286.68	284.62	216.25	307.80	309.70	339.02	295.48	156.93	57.34
Lingua	I_s	554.94	553.15	607.40	617.07	567.67	665.73	645.21	743.76	643.01	325.39	110.21
	I_l	263.89	258.09	292.36	286.70	280.85	317.55	312.28	346.24	**318.18**	163.83	50.08
Ours	I_s	**600.62**	**611.43**	**625.14**	**648.91**	**587.98**	**677.77**	**678.51**	**756.44**	**648.90**	**356.55**	104.32
	I_l	**283.54**	**296.09**	**298.73**	**308.92**	**292.74**	**332.16**	**326.67**	**357.74**	318.06	**191.09**	**44.33**
Δ	I_s	+47.30	+60.64	+40.02	+52.60	+12.94	+12.85	+94.94	+55.08	+73.90	+104.56	30.32
	I_l	+21.47	+44.05	+19.87	+26.29	+8.70	+13.79	+33.25	+20.60	+14.76	+89.76	23.57

Table 3: The AUC ↑ results for the EntityQuestions dataset. The symbol definitions are same as Table 2.

D	K	G-1.3	G-2.7	O-1.3	O-2.7	b-560	b-7b1	L-13	L3.1-8	DS-V2	Q3-32	$\sigma \downarrow$
Vanilla	I_s	**550.08**	608.54	618.05	**677.63**	**511.98**	**705.35**	657.06	743.99	572.72	235.42	142.98
	I_l	**259.35**	283.86	284.91	**318.26**	**236.82**	**329.58**	296.63	338.60	**313.36**	87.65	72.88
TF-IDF	I_s	302.59	459.72	419.50	517.23	314.45	552.43	666.08	627.44	180.75	235.64	165.91
	I_l	146.52	239.16	188.60	259.91	155.99	273.13	323.23	276.02	107.46	112.64	75.92
Keywords	I_s	358.34	458.67	495.48	545.41	392.71	572.18	614.18	674.23	284.15	287.12	135.78
	I_l	171.09	229.08	245.89	276.19	190.40	282.74	310.78	323.42	175.65	128.99	65.40
Summary	I_s	336.92	366.90	450.84	437.40	396.18	498.25	435.01	511.30	448.16	210.08	**88.12**
	I_l	161.38	180.04	221.94	202.50	196.11	254.38	209.77	242.42	247.76	77.62	**52.17**
SelCon	I_s	278.08	329.18	359.08	391.45	251.39	401.26	531.96	545.13	395.29	226.52	107.42
	I_l	136.32	163.02	177.21	187.91	137.72	195.78	268.26	259.44	208.08	103.98	52.52
Lingua	I_s	541.93	598.45	592.69	644.01	496.46	670.92	698.64	792.93	648.58	374.74	115.86
	I_l	244.38	275.40	274.64	283.11	223.05	308.36	322.57	357.82	307.12	152.43	57.73
Ours	I_s	546.46	**627.41**	**632.79**	662.16	494.45	688.73	**738.82**	**813.86**	**652.14**	**406.00**	118.33
	I_l	248.82	**294.48**	**298.31**	295.18	229.06	323.26	**343.58**	**371.30**	307.05	**181.50**	55.95
Δ	I_s	-3.62	+18.87	+14.74	-15.47	-17.53	-16.62	+81.76	+69.87	+79.42	+170.58	61.27
	I_l	-10.53	+10.62	+13.40	-23.08	-7.76	-6.32	+46.95	+32.70	-6.31	+93.85	35.11

sentation showing lower sensitivity to surface-level changes. In contrast, the discrete keywords-based compression shows notable performance swings. These observations answer **Q2** by showing that LLM-driven baselines are not a reliable choice due to the uncertainty in inference.

Compared with the SelCon baseline, our method achieves higher AUC across configurations. We hypothesize that this gap stems from fundamental differences in our approaches: while both methods utilize information theory, SelCon operates at the phrase/sentence level through token-based self-information aggregation for content filtering, whereas our method uses AMR's structured semantic representation to compute concept-level entropy based on semantic roles and connections in comprehensive contexts. The AMR-based entropy better preserves the conceptual coherence for complex reasoning, as it captures semantic structures and dependencies that are crucial for maintaining clear inferential chains for reconstructing scenarios.

LLMLingua represents competitive baseline as a token-level compression technique. The advantage of our method relative to LLMLingua comes from the complementary strengths of semantic-level versus token-level compression: while LLMLingua selects tokens through iterative perplexity-based filtering and budget control, our AMR-based approach identifies coherent concept units that match the information structure. Both methods preserve essential information, but our semantic abstraction excels when maintaining conceptual relationships matters more than surface-level linguistic continuity. Moreover, our method enhances the interpretability and readability by preserving complete conceptual units as atomic elements and maintain-

ing lexical integrity, whereas token-level compression can fragment words that disrupt local linguistic structures. This property facilitates human understanding and debugging. Compared with other baselines, both SelCon and LLMLingua achieve competitive AUC and σ, addressing **Q3** on the necessity of dedicated context compression methods.

5.2 Performance on Long Contexts

To further validate our method and highlight its characteristics, we analyze performance in the long-context interval I_l in Table 2 and Table 3, emphasizing behaviors that emerge specifically under long-context conditions. The proposed method achieves the competitive performance that keeps the same trend as in the I_s, but the gains are reduced. The reduction is expected since the I_l interval typically encompasses longer contexts or higher complexity scenarios, where the marginal benefit of improvements tends to diminish. However, a notable phenomenon is that σ is significantly lower for this interval, which contains longer but more concentrated concepts compared with the massive but dispersed interval, indicating the benefit of macro-level semantic constraints in capturing informative concepts within complex contexts in specific scenarios. Moreover, the low σ of Δ indicates consistent performance variance across backbones.

5.3 Compression Efficiency

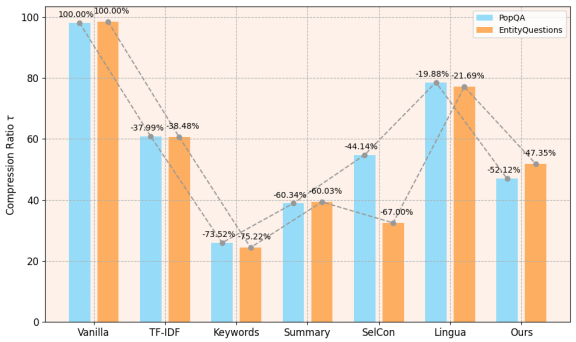

Figure 3: Comparison of token-level compression ratios across different context compression methods.

We examine the compression efficiency in terms of token-level reduction (τ) and inference latency (ms per instance). As shown in Figure 3, our method reduces the length to about 50% of the vanilla on average, while keeping the Acc stable in both datasets. Baselines such as Keywords and Summary yield lower token counts, but they often remove meaningful factual cues, leading to performance drops. In contrast, operating at the concept level through AMR allows the compressed context to retain the core semantic units needed for reasoning, rather than relying on surface lexical signals.

Table 4: Inference time comparison (ms per instance)

LLMs	Vanilla	TF-IDF	Keywords	Summary	SelCon	Lingua	Ours
PopQA							
G-1.3	402.89	468.01	**366.23**	410.52	470.27	429.32	380.38
G-2.7	672.12	622.81	**548.13**	578.64	634.32	640.69	548.51
O-1.3	322.68	314.18	**281.84**	316.45	305.92	356.40	306.23
O-2.7	517.73	499.23	484.59	487.71	524.98	526.04	**461.01**
b-560	261.43	265.57	**235.13**	237.32	275.10	274.51	249.55
b-7b1	1130.13	1152.86	**1006.23**	1006.60	1150.33	1139.21	1058.83
L-13	1886.29	1405.44	**1329.71**	1364.58	1476.61	1507.88	1409.22
L3.1-8	1032.17	1091.39	688.09	**644.89**	1109.62	1089.12	888.58
DS-V2	1233.80	166.51	**150.13**	165.69	293.25	171.14	164.05
Q3-32	5283.06	5029.57	**4795.76**	4879.41	5094.38	5040.01	4783.34
EntityQuestions							
G-1.3	605.82	587.49	**546.79**	547.65	724.10	761.94	585.63
G-2.7	866.79	811.66	749.83	**746.46**	867.93	932.25	779.43
O-1.3	528.14	**486.72**	481.75	496.73	557.98	648.45	499.03
O-2.7	703.28	684.91	**647.43**	671.47	761.57	827.20	702.26
b-560	445.16	468.14	421.71	**416.70**	527.12	582.26	439.89
b-7b1	1319.82	1338.23	1196.88	**1176.98**	1190.92	1456.20	1279.33
L-13	1805.86	1786.85	1672.17	1743.26	1717.31	1881.69	**1590.65**
L3.1-8	1233.70	1282.96	871.17	**836.18**	1016.64	1398.92	1083.90
DS-V2	358.03	**326.23**	333.82	326.80	431.81	444.87	330.10
Q3-32	5239.69	5313.41	**4996.61**	5012.55	5120.52	5409.85	5168.47

The reduction in context length leads directly to faster inference, and the latency decreases in line with the length reduction. Table 4 shows that the proposed method lowers the average inference time compared to the vanilla setting. Baselines reducing latency via token pruning may fragment expressions and weaken local coherence, especially in long contexts. By retaining intact conceptual units, our compressed contexts remain stable for reasoning, enabling both shorter inference time and reliable answering, even under high compression.

6 Conclusion

This paper presents a compression method for context engineering that leverages conceptual information entropy of AMR to identify semantically crucial concepts. Our method shows improvements over baselines while achieving substantial compression ratios. The experiments demonstrate that AMR-based semantic analysis guides context compression effectively. The integration of structured linguistic representation with information-theoretic concept selection offers a paradigm to balance information retention with computational efficiency.

Future research includes extending our approach to multi-modal contexts, modeling cross-document concept relationships, and exploring adaptive compression strategies based on query complexity. Incorporating other stable linguistic representations is also a valuable direction to improve the efficiency and effectiveness in context engineering.

Limitations

Although the proposed method shows clear gains in long-context settings, some limitations remain. First, the current approach relies on the stability of AMR parsers, and the performance may decline when the parser produces incomplete or noisy graphs. The parsing processing is based on the sentence-level graph, so complex document-level structures are easily ignored. These dependency introduces upper bounds on covered conceptual information in compression. Developing reliable AMR parsers is a continuously valuable direction.

Second, the current setup evaluates compression under a controlled testing environment where answer-containing documents are considered. This design isolates the effect of compression but does not fully reflect real-world retrieval pipelines, where irrelevant or conflicting documents are common. Experimenting with the setting in a full retrieval stack and examining different retrievers' influence will be conducted in future work.

Finally, computing AMR graphs and entropy scores introduces extra cost during preprocessing. Although this cost occurs offline, it may restrict the method in latency-sensitive systems or in large-scale applications where many documents must be processed. A crucial future work is exploring high-efficiency solutions for these stages.

References

Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. 2025. The unreasonable effectiveness of entropy minimization in llm reasoning. arXiv preprint arXiv:2505.15134.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Michele Bevilacqua, Rexhina Blloshmi, and Roberto Navigli. 2021. One spring to rule them both: Symmetric amr semantic parsing and generation without a complex pipeline. In Proceedings of the AAAI conference on artificial intelligence, volume 35, pages 12564–12573.

Jeffrey R Binder, Rutvik H Desai, William W Graves, and Lisa L Conant. 2009. Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies. Cerebral cortex, 19(12):2767–2796.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In Advances in Neural Information Processing Systems, volume 33, pages 1877–1901. Curran Associates, Inc.

Qi Cao, Takeshi Kojima, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Unnatural error correction: GPT-4 can almost perfectly handle unnatural scrambled text. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 8898–8913, Singapore. Association for Computational Linguistics.

Huiyao Chen, Meishan Zhang, Jing Li, Min Zhang, Lilja Øvrelid, Jan Hajič, and Hao Fei. 2025. Semantic role labeling: A systematical survey. arXiv preprint arXiv:2502.08660.

Xin Cheng, Xun Wang, Xingxing Zhang, Tao Ge, Si-Qing Chen, Furu Wei, Huishuai Zhang, and Dongyan Zhao. 2024. xrag: Extreme context compression for retrieval-augmented generation with one token. Advances in Neural Information Processing Systems, 37:109487–109516.

DeepSeek-AI. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. Preprint, arXiv:2405.04434.

Changde Du, Kaicheng Fu, Bincheng Wen, Yi Sun, Jie Peng, Wei Wei, Ying Gao, Shengpei Wang, Chuncheng Zhang, Jinpeng Li, and 1 others. 2025. Human-like object concept representations emerge naturally in multimodal large language models. Nature Machine Intelligence, pages 1–16.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. CoRR, abs/2407.21783.

Evelina Fedorenko, Steven T Piantadosi, and Edward AF Gibson. 2024. Language is primarily a tool for communication rather than thought. Nature, 630(8017):575–586.

Tomoyasu Horikawa. 2025. Mind captioning: Evolving descriptive text of mental content from human brain activity. Science Advances, 11(45):eadw1464.

Taeho Hwang, Sukmin Cho, Soyeong Jeong, Hoyun Song, SeungYoon Han, and Jong C Park. 2024. Exit: Context-aware extractive compression for enhancing retrieval-augmented generation. arXiv preprint arXiv:2412.12559.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. Transactions on Machine Learning Research.

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, pages 13358–13376, Singapore. Association for Computational Linguistics.

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 (ACL), pages 1658–1677.

Jiajie Jin, Xiaoxi Li, Guanting Dong, Yuyao Zhang, Yutao Zhu, Yongkang Wu, Zhonghua Li, Qi Ye, and Zhicheng Dou. 2025a. Hierarchical document refinement for long-context retrieval-augmented generation. arXiv preprint arXiv:2505.10413.

Yiqiao Jin, Kartik Sharma, Vineeth Rakesh, Yingtong Dou, Menghai Pan, Mahashweta Das, and Srijan Kumar. 2025b. Sara: Selective and adaptive retrieval-augmented generation with context compression. arXiv preprint arXiv:2507.05633.

Zhijing Jin, Yuen Chen, Fernando Gonzalez Adauto, Jiarui Liu, Jiayi Zhang, Julian Michael, Bernhard Schölkopf, and Mona Diab. 2024. Analyzing the role of semantic representations in the era of large language models. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 3781–3798, Mexico City, Mexico. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6769–6781. Association for Computational Linguistics.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In The Eleventh International Conference on Learning Representations.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, and 1 others. 2022. Bloom: A 176b-parameter open-access multilingual language model.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Igor Kulikov, Vishrav Chaudhary, Sebastian Wang, Wentau Yih Barta, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In Advances in Neural Information Processing Systems, volume 33, pages 9459–9474.

Chenliang Li, Bin Bi, Ming Yan, Wei Wang, and Songfang Huang. 2021. Addressing semantic drift in generative question answering with auxiliary extraction. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 942–947, Online. Association for Computational Linguistics.

Yucheng Li, Bo Dong, Frank Guerin, and Chenghua Lin. 2023a. Compressing context to enhance inference efficiency of large language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 6342–6353, Singapore. Association for Computational Linguistics.

Yucheng Li, Bo Dong, Chenghua Lin, and Frank Guerin. 2023b. Compressing context to enhance inference efficiency of large language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Fei Liu, Jeffrey Flanigan, Sam Thomson, Norman Sadeh, and Noah A. Smith. 2015. Toward abstractive summarization using semantic representations. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1077–1086, Denver, Colorado. Association for Computational Linguistics.

Shengjie Liu, Jing Wu, Jingyuan Bao, Wenyi Wang, Naira Hovakimyan, and Christopher G Healey. 2024. Towards a robust retrieval-based summarization system. arXiv preprint arXiv:2403.19889.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Wenhao Mao, Chengbin Hou, Tianyu Zhang, Xinyu Lin, Ke Tang, and Hairong Lv. 2024. Parse trees guided llm prompt compression. arXiv preprint arXiv:2409.15395.

Lingrui Mei, Jiayu Yao, Yuyao Ge, Yiwei Wang, Baolong Bi, Yujun Cai, Jiazhi Liu, Mingyu Li, Zhong-Zhi Li, Duzhen Zhang, and 1 others. 2025. A survey of context engineering for large language models. arXiv preprint arXiv:2507.13334.

Dang Nguyen, Ali Payani, and Baharan Mirzasoleiman. 2025. Beyond semantic entropy: Boosting LLM uncertainty quantification with pairwise semantic similarity. In Findings of the Association for Computational Linguistics: ACL 2025, pages 4530–4540, Vienna, Austria. Association for Computational Linguistics.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. 2024. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. In Advances in Neural Information Processing Systems, volume 37, pages 8901–8929. Curran Associates, Inc.

Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. Foundations and Trends® in Information Retrieval, 3(4):333–389.

Timothy T Rogers, Matthew A Lambon Ralph, Peter Garrard, Sasha Bozeat, James L McClelland, John R Hodges, and Karalyn Patterson. 2004. Structure and deterioration of semantic memory: a neuropsychological and computational investigation. Psychological review, 111(1):205.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple entity-centric questions challenge dense retrievers. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6138–6148, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chunting Shi, Michihiro Yasunaga, Isabelle Augenstein, Nikos Voskarides, Mikel Artetxe, Xiang Ren, Xiaozhong Wan, Antoine Bosselut, Dragomir Radev, Wenpeng Yin, and 1 others. 2023. Replug: Retrieval-augmented black-box language models. arXiv preprint arXiv:2301.12652.

Kaize Shi, Xueyao Sun, Qing Li, and Guandong Xu. 2024. Compressing long context for enhancing rag with amr-based concept distillation. arXiv preprint arXiv:2405.03085.

Linfeng Song, Daniel Gildea, Yue Zhang, Zhiguo Wang, and Jinsong Su. 2019. Semantic neural machine translation using amr. Transactions of the Association for Computational Linguistics, 7:19–31.

Siddharth Suresh, Kushin Mukherjee, Xizheng Yu, Wei-Chun Huang, Lisa Padua, and Timothy Rogers. 2023. Conceptual structure coheres in human cognition but not in large language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 722–738, Singapore. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Qwen Team. 2025. Qwen3 technical report. Preprint, arXiv:2505.09388.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Sourav Verma. 2024. Contextual compression in retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2409.13385.

Jingyao Wang, Wenwen Qiang, Zeen Song, Changwen Zheng, and Hui Xiong. 2025. Learning to think: Information-theoretic reinforcement fine-tuning for llms. arXiv preprint arXiv:2505.10425.

Shira Wein and Juri Opitz. 2024. A survey of AMR applications. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 6856–6875, Miami, Florida, USA. Association for Computational Linguistics.

EC Wit and Marie Gillette. 1999. What is linguistic redundancy. University of Chicago.

Kevin Wu, Eric Wu, and James Zou. 2024. Clasheval: Quantifying the tug-of-war between an LLM's internal prior and external evidence. In The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track.

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

Qihui Xu, Yingying Peng, Samuel A Nastase, Martin Chodorow, Minghua Wu, and Ping Li. 2025. Large language models without grounding recover non-sensorimotor but not sensorimotor features of human concepts. Nature human behaviour, pages 1–16.

Mingjia Yin, Chuhan Wu, Yufei Wang, Hao Wang, Wei Guo, Yasheng Wang, Yong Liu, Ruiming Tang, Defu Lian, and Enhong Chen. 2024. Entropy law: The story behind data compression and llm performance. arXiv preprint arXiv:2407.06645.

Jiahuan Zhang, Tianheng Wang, Hanqing Wu, Ziyi Huang, Yulong Wu, Dongbai Chen, Linfeng Song, Yue Zhang, Guozheng Rao, and Kaicheng Yu. 2025. Sr-llm: Rethinking the structured representation in large language model. arXiv preprint arXiv:2502.14352.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, and 1 others. 2022. Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068.

Chuyue Zhou, Wangjie You, Juntao Li, Jing Ye, Kehai Chen, and Min Zhang. 2023. INFORM : Information eNtropy based multi-step reasoning FOR large language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 3565–3576, Singapore. Association for Computational Linguistics.

Jiawei Zhou, Tahira Naseem, Ramón Fernandez Astudillo, and Radu Florian. 2021. AMR parsing with action-pointer transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5585–5598, Online. Association for Computational Linguistics.

A Prompts for Baselines

Following the instruction-tuning framework of Taori et al. (2023), we design prompt templates for keyword extraction and summarization baselines, as detailed in Prompt A1 and Prompt A2.

Prompt A1: Keywords Extraction

[INST] «SYS»
`Extract a few keywords from the following content.`
«/SYS»
Prompt = """Below is an instruction that describes a task, paired with an input that provides content.
Instruction: {""" + `Instruction` + """}
Input: {""" + D + """}
Response: """
[/INST]

Prompt A2: Summary Generation

[INST] «SYS»
`Generate a short summary of the following content.`
«/SYS»
Prompt = """Below is an instruction that describes a task, paired with an input that provides content.
Instruction: {""" + `Instruction` + """}
Input: {""" + D + """}
Response: """
[/INST]

B Accuracy Details

C Ablation Study

We perform an ablation study to analyze the impact of the hyper-parameter α, which controls the significance threshold in the concept pruning process, on the overall performance of our method and the results are shown in Table A3. This parameter determines which concepts are retained from the AMR graphs based on their entropy values to construct the compressed context. Table A3 shows that the lower values of α overly restrict the retained information, pruning out useful concepts and leading to degraded performance. Conversely, higher α values retain too many concepts, which may introduce noise and reduce compression efficiency.

Based on the aforementioned observation, we set $\alpha = 0.3$ in our method, which represents the optimal trade-off, maximizing the discriminatory power of retained concepts while maintaining a compact and informative context for downstream inference. This tuning contributes significantly to the robustness and effectiveness of our context compression approach in context engineering.

Table A1: Accuracy (Acc ↑) comparison on the PopQA dataset. The best results for each LLM with setting K are in **bold**, and the next best results are in underlined. Δ here represents the difference between Ours and Vanilla, and the increased and decreased Δ are marked differently. The best results for each of K are marked.

LLMs	C \ K	1	2	3	4	5	6	7	8	9	10
GPT-Neo-1.3B	Vanilla	48.57	64.77	54.60	54.07	63.13	60.78	62.42	63.23	**69.63**	72.80
	TF-IDF	22.86	32.89	38.51	38.95	42.50	39.87	42.28	38.71	43.70	50.40
	Keywords	30.00	41.61	43.68	50.58	53.75	50.98	48.32	47.10	43.70	57.60
	Summary	23.93	41.61	47.13	48.84	52.50	50.33	46.98	48.39	56.30	58.40
	SelCon	39.29	54.03	42.53	48.26	51.88	55.56	46.98	49.03	52.59	65.60
	LLMLingua	**53.93**	53.69	56.90	58.72	64.38	60.78	63.76	65.81	65.93	76.00
	Ours	**53.93**	**68.45**	**57.47**	**59.88**	**70.00**	68.63	**69.13**	**71.61**	68.89	**79.20**
	$\alpha = 0.01$	19.64	32.21	42.53	48.84	53.13	54.90	65.10	67.10	62.96	76.80
	$\alpha = 0.05$	28.57	41.28	48.28	56.98	58.75	60.78	67.11	68.39	63.70	77.60
	$\alpha = 0.1$	28.57	41.28	48.28	56.98	58.75	60.78	67.11	68.39	63.70	77.60
	$\alpha = 0.5$	35.00	51.01	56.90	58.14	61.88	**69.28**	**69.13**	62.58	65.19	76.80
	Δ	+5.36	+3.68	+2.87	+5.81	+6.87	+7.85	+6.71	+8.38	-0.74	+6.40
GPT-Neo-2.7B	Vanilla	51.07	69.46	59.77	52.91	59.38	63.40	59.73	57.42	65.19	76.00
	TF-IDF	33.57	46.31	50.00	53.49	58.75	64.05	59.06	61.29	60.74	76.00
	Keywords	30.71	43.00	48.28	51.16	54.38	52.29	56.38	47.10	51.85	59.20
	Summary	22.86	39.93	50.00	50.00	56.25	56.21	52.35	57.42	55.56	60.80
	SelCon	45.00	56.71	50.00	46.51	59.38	54.25	57.72	54.84	53.33	70.40
	LLMLingua	**52.14**	70.13	59.20	50.54	59.38	59.48	63.76	61.29	63.70	79.20
	Ours	51.07	**70.45**	60.34	61.63	**64.38**	**66.01**	**75.17**	69.68	**77.03**	**82.40**
	$\alpha = 0.01$	20.71	32.21	40.23	55.23	53.13	60.12	59.06	64.52	62.96	77.60
	$\alpha = 0.05$	29.64	41.61	53.45	56.98	60.63	63.40	68.46	64.52	65.93	76.00
	$\alpha = 0.1$	30.36	47.99	50.57	59.30	62.50	64.05	72.48	**71.61**	71.11	81.60
	$\alpha = 0.5$	40.36	53.02	**61.49**	**64.53**	**64.38**	**66.01**	72.48	67.74	69.63	81.60
	Δ	0.00	+0.99	+0.57	+8.72	+5.00	+2.61	+15.44	+12.26	+11.84	+6.40
OPT-1.3b	Vanilla	52.14	67.11	63.22	57.56	61.25	62.09	69.13	71.62	66.67	80.80
	TF-IDF	36.79	47.32	46.55	47.09	53.75	58.17	59.06	60.00	61.48	68.80
	Keywords	31.79	45.30	55.17	55.23	64.38	64.05	65.10	68.39	60.74	76.80
	Summary	26.79	47.65	56.90	59.30	65.00	61.44	65.10	67.74	65.19	77.60
	SelCon	49.29	61.07	56.90	56.98	63.75	60.13	67.79	73.55	74.07	82.40
	LLMLingua	**55.00**	**71.14**	61.49	57.56	65.00	64.71	**75.17**	73.54	68.89	84.80
	Ours	54.29	69.13	**68.39**	59.30	**68.13**	68.62	74.50	74.19	73.33	84.80
	$\alpha = 0.01$	23.57	34.56	44.25	48.84	58.13	60.13	69.80	73.55	70.37	84.80
	$\alpha = 0.05$	30.36	44.30	55.17	59.88	60.00	62.09	73.83	73.55	70.37	82.40
	$\alpha = 0.1$	32.86	46.98	60.92	**62.79**	66.25	**69.93**	73.15	75.48	**75.56**	**86.40**
	$\alpha = 0.5$	42.14	57.72	60.34	59.88	65.63	68.28	**75.17**	**77.42**	71.85	84.00
	Δ	+2.15	+2.02	+5.17	+1.74	+6.88	+6.53	+5.37	+2.57	+6.66	+4.00
OPT-2.7b	Vanilla	49.64	66.78	62.64	63.72	65.00	61.44	64.43	70.32	75.56	83.20
	TF-IDF	33.21	48.32	49.43	51.16	56.88	64.71	64.43	70.32	65.19	73.60
	Keywords	35.36	43.96	58.05	58.14	65.00	59.48	66.44	70.97	68.89	76.80
	Summary	29.64	49.33	54.60	55.23	60.00	54.90	60.40	65.16	56.30	67.20
	SelCon	48.21	64.09	54.02	58.14	66.25	60.78	67.11	74.19	73.33	79.20
	LLMLingua	**55.71**	**73.15**	62.64	62.79	71.25	65.36	63.09	77.42	71.11	**84.80**
	Ours	55.36	70.13	**64.37**	**70.35**	**72.50**	69.93	**76.51**	**78.06**	77.78	83.20
	$\alpha = 0.01$	22.86	35.91	45.98	54.65	61.88	62.75	66.44	72.26	68.15	83.20
	$\alpha = 0.05$	33.57	45.30	59.77	61.05	66.25	66.67	73.15	**78.06**	77.78	80.80
	$\alpha = 0.1$	35.00	53.02	60.34	66.86	69.38	67.32	75.17	76.13	**79.26**	80.00
	$\alpha = 0.5$	45.36	63.42	62.64	68.02	66.25	**73.20**	73.83	71.61	77.78	80.00
	Δ	+5.72	+3.35	+1.73	+6.63	+7.50	+8.49	+12.08	+7.74	+2.22	0.00
Bloom-560m	Vanilla	51.07	62.42	54.02	**56.40**	61.25	62.75	66.44	72.90	**73.33**	80.00
	TF-IDF	27.14	34.90	36.78	43.60	48.75	45.10	57.72	52.26	52.59	64.80
	Keywords	26.43	44.30	**56.32**	48.26	55.63	56.21	62.42	63.23	60.74	75.20
	Summary	27.50	47.65	52.87	54.07	61.88	58.17	67.11	70.97	62.22	77.60
	SelCon	34.29	49.66	45.40	43.60	47.50	47.06	51.68	59.35	48.89	65.60
	LLMLingua	**53.57**	65.10	52.87	52.33	60.00	59.48	68.46	74.84	67.41	**80.80**
	Ours	52.86	**66.44**	55.74	55.23	59.38	**64.05**	**71.81**	**76.77**	**73.33**	77.60
	$\alpha = 0.01$	18.57	26.51	33.91	36.63	46.25	50.33	61.07	60.65	66.67	72.80
	$\alpha = 0.05$	26.07	39.93	41.95	49.42	51.88	54.25	70.47	61.29	68.89	71.20
	$\alpha = 0.1$	30.00	40.94	46.55	47.09	50.63	61.44	71.14	65.16	71.85	76.80
	$\alpha = 0.5$	33.21	44.63	54.02	53.49	**63.13**	62.75	71.14	69.03	63.70	72.00
	Δ	+1.79	+4.02	+1.72	-1.17	-1.87	+1.30	+5.37	+3.87	0.00	-2.40
Bloom-7b1	Vanilla	56.43	**73.49**	72.41	65.12	68.75	**77.12**	78.52	80.65	77.04	87.20
	TF-IDF	40.36	56.04	62.64	61.04	65.63	71.24	76.51	76.13	77.04	84.80
	Keywords	38.93	53.02	62.64	59.88	65.63	67.32	75.84	74.19	71.85	77.60
	Summary	32.50	49.66	60.34	56.40	64.38	71.90	74.50	71.61	74.07	77.60
	SelCon	53.21	66.11	67.24	65.70	65.00	71.90	74.50	78.71	77.04	83.20
	LLMLingua	**57.86**	72.82	72.99	**67.44**	68.75	74.50	75.84	81.94	78.52	88.00
	Ours	54.64	69.80	**73.56**	65.12	71.25	**77.12**	**82.55**	**83.23**	82.22	91.20
	$\alpha = 0.01$	22.14	35.57	43.68	50.58	64.38	61.44	71.14	70.97	71.85	82.40
	$\alpha = 0.05$	31.07	48.66	59.77	59.88	68.13	68.63	73.83	74.84	80.74	86.40
	$\alpha = 0.1$	36.43	51.68	59.20	59.88	70.63	71.90	75.17	77.42	**82.96**	89.60
	$\alpha = 0.5$	42.86	61.74	66.67	62.79	**73.75**	75.82	79.19	82.58	81.48	85.60
	Δ	-1.79	-3.69	+1.15	0.00	+2.50	0.00	+4.05	+2.58	+5.18	+4.00
Llama-2-chat-13b	Vanilla	51.78	60.40	56.32	61.04	59.38	54.24	69.80	74.84	79.26	84.80
	TF-IDF	48.57	59.01	63.79	62.21	65.63	73.20	77.18	78.71	77.78	82.40
	Keywords	36.43	52.01	55.17	58.72	58.13	62.75	68.46	72.26	69.63	74.40
	Summary	27.86	44.63	46.55	54.07	48.13	46.41	56.38	65.16	65.93	83.20
	SelCon	55.00	69.13	63.79	67.44	65.00	69.28	73.15	78.71	80.00	86.40
	LLMLingua	58.93	68.79	63.22	71.51	65.63	68.63	72.48	80.00	**81.48**	88.00
	Ours	**59.64**	**69.46**	**69.54**	**72.67**	**73.75**	73.20	**81.21**	**82.58**	**81.48**	**89.60**
	$\alpha = 0.01$	30.00	41.61	44.83	54.65	56.25	66.01	71.14	68.39	73.33	85.60
	$\alpha = 0.05$	36.79	53.69	55.74	56.40	62.50	64.71	73.83	72.26	76.30	83.20
	$\alpha = 0.1$	43.93	61.41	62.64	63.37	65.63	70.59	75.17	75.48	77.03	84.80
	$\alpha = 0.5$	55.00	68.46	68.97	69.19	70.00	**77.12**	80.54	78.06	77.04	82.40
	Δ	+7.86	+9.06	+13.22	+11.63	+14.37	+18.96	+11.41	+7.74	+2.22	+4.80
Llama-3.1-8B-Instruct	Vanilla	61.43	76.17	72.41	74.42	70.63	79.74	83.22	82.58	86.67	89.60
	TF-IDF	51.07	66.78	64.37	70.93	70.63	73.20	78.52	81.94	81.48	75.20
	Keywords	51.79	56.04	62.07	63.37	61.88	69.28	73.83	76.13	77.04	81.60
	Summary	38.93	56.38	55.75	65.12	50.63	67.32	68.46	67.10	65.19	72.00
	SelCon	68.21	76.85	76.44	76.16	75.63	79.08	84.56	87.10	82.22	91.20
	LLMLingua	73.21	85.23	78.74	77.91	76.88	84.31	83.89	86.45	88.15	91.20
	Ours	73.93	80.20	81.61	79.65	78.13	84.31	89.93	89.68	90.37	91.20
	$\alpha = 0.01$	40.71	56.71	57.47	68.02	66.88	75.82	85.23	84.52	86.67	89.60
	$\alpha = 0.05$	50.36	67.45	67.24	74.42	73.75	80.39	87.25	85.16	88.15	88.00
	$\alpha = 0.1$	55.00	70.81	72.99	77.91	78.13	84.97	89.26	85.81	89.63	88.80
	$\alpha = 0.5$	67.50	80.54	79.31	79.07	78.13	83.66	89.26	85.16	86.67	89.60
	Δ	+12.50	+4.03	+9.20	+5.23	+7.50	+4.57	+6.71	+7.10	+3.70	+1.60
DeepSeek-V2-Lite	Vanilla	17.86	49.33	53.45	65.70	60.63	67.32	71.15	80.65	77.04	81.60
	TF-IDF	7.14	7.38	11.49	14.53	23.75	24.18	18.79	27.74	31.85	32.00
	Keywords	6.07	16.78	21.26	24.42	35.00	42.48	40.94	41.94	42.96	53.60
	Summary	9.64	35.91	44.25	52.91	53.13	62.09	57.72	67.74	72.59	80.80
	SelCon	23.93	54.70	49.43	56.98	57.50	62.75	72.48	78.06	75.56	76.00
	LLMLingua	35.00	**67.79**	67.24	**72.09**	66.88	66.67	**75.17**	**83.23**	84.44	84.00
	Ours	40.36	67.45	**68.39**	**72.09**	**68.75**	**67.97**	73.83	80.65	**85.19**	**88.80**
	$\alpha = 0.01$	23.21	37.92	35.63	47.67	41.88	47.06	58.39	56.77	59.26	68.00
	$\alpha = 0.05$	27.50	44.97	41.38	50.58	41.88	47.06	55.03	57.42	64.44	56.80
	$\alpha = 0.1$	33.21	48.32	45.98	49.42	46.88	56.21	57.72	59.35	65.93	76.80
	$\alpha = 0.5$	**43.57**	55.70	47.13	54.65	53.13	63.40	69.80	76.13	79.26	78.40
	Δ	+22.50	+18.12	+14.94	+6.39	+8.12	+0.65	+2.68	0.00	+8.15	+7.20
Qwen3-32B	Vanilla	27.50	29.53	32.76	34.88	24.38	30.72	21.48	28.39	23.70	24.80
	TF-IDF	14.29	14.77	17.82	21.51	26.25	18.30	23.49	29.03	36.30	32.00
	Keywords	22.86	26.85	22.99	27.33	27.50	28.10	34.23	38.06	32.59	45.60
	Summary	22.86	28.52	30.46	29.65	29.38	33.99	28.19	35.48	34.07	48.00
	SelCon	21.07	26.51	26.44	26.74	31.25	29.41	42.95	34.19	41.48	47.20
	LLMLingua	**30.71**	**30.54**	**36.78**	36.05	28.13	29.41	32.21	44.52	45.19	54.40
	Ours	28.92	**30.54**	36.20	**36.62**	30.00	35.29	**43.62**	45.81	**54.81**	**58.40**
	$\alpha = 0.01$	11.79	19.80	25.86	24.42	**33.13**	29.41	37.58	41.29	37.78	40.80
	$\alpha = 0.05$	15.36	25.50	29.31	27.33	31.88	31.37	40.27	**52.26**	**54.81**	55.20
	$\alpha = 0.1$	18.21	29.87	31.61	30.81	28.75	**35.95**	**43.62**	42.58	52.59	51.20
	$\alpha = 0.5$	25.36	27.18	33.91	34.88	29.38	30.72	36.91	34.84	40.00	40.80
	Δ	+1.42	+1.01	+3.44	+1.74	+5.62	+4.57	+22.14	+17.42	+31.11	+33.60

Table A2: Accuracy (Acc ↑) comparison the EntityQuestions dataset. The symbols' definitions are same as Table A1.

LLMs	C' \ K	1	2	3	4	5	6	7	8	9	10
GPT-Neo-1.3B	Vanilla	47.24	60.31	58.45	56.95	60.25	62.31	60.34	65.09	**66.92**	**71.68**
	TF-IDF	21.27	28.32	32.71	29.15	35.15	40.20	33.52	37.28	36.15	38.94
	Keywords	22.50	35.66	37.27	36.61	43.10	46.73	46.93	44.38	33.85	45.13
	Summary	22.29	34.09	36.46	35.93	41.84	32.16	36.87	40.83	43.85	47.49
	SelCon	21.06	25.70	29.76	29.15	31.80	29.65	30.17	35.51	40.77	30.09
	LLMLingua	**51.53**	**64.16**	**60.86**	56.95	56.90	**65.83**	62.01	58.58	57.69	66.37
	Ours	50.92	61.36	59.79	57.29	**64.85**	57.79	60.35	63.31	63.08	66.37
	$\alpha = 0.01$	19.02	30.59	40.48	42.37	40.17	44.72	55.31	53.25	53.08	58.41
	$\alpha = 0.05$	25.97	39.69	44.77	49.49	52.72	53.27	59.78	63.91	55.38	59.29
	$\alpha = 0.1$	28.63	46.33	50.94	53.56	58.58	59.30	**63.69**	**66.27**	58.46	60.18
	$\alpha = 0.5$	37.01	51.05	54.69	**57.63**	55.23	56.78	55.87	56.21	56.15	59.29
	Δ	+3.68	+1.05	+1.34	+0.34	+4.60	-4.52	+0.01	-1.78	-3.84	-5.31
GPT-Neo-2.7B	Vanilla	**54.40**	64.86	65.42	64.75	67.78	69.35	71.51	68.64	72.31	73.45
	TF-IDF	30.88	33.39	49.33	43.73	51.04	55.28	60.34	57.99	60.00	66.37
	Keywords	29.65	41.78	46.92	48.14	49.79	56.28	55.87	59.17	58.46	54.87
	Summary	21.06	35.14	39.14	35.93	42.26	47.74	39.66	44.38	50.00	44.25
	SelCon	24.74	28.67	35.92	32.20	37.66	38.69	44.13	41.42	43.08	30.09
	LLMLingua	54.19	**65.56**	63.81	62.71	**70.71**	66.33	70.95	68.64	67.69	69.91
	Ours	54.21	62.94	**69.71**	66.78	**70.71**	**71.36**	**74.86**	71.60	**73.85**	**76.99**
	$\alpha = 0.01$	20.45	33.22	48.53	48.47	50.21	50.75	69.27	62.13	60.77	69.91
	$\alpha = 0.05$	30.67	44.76	57.64	60.00	58.16	62.81	68.72	71.60	63.85	72.57
	$\alpha = 0.1$	36.20	52.27	59.59	62.71	63.60	63.82	72.07	71.01	67.69	68.14
	$\alpha = 0.5$	48.46	60.14	68.10	**68.14**	69.04	70.35	71.51	**73.96**	71.54	73.45
	Δ	-0.19	-1.92	+4.29	+2.03	+2.93	+2.01	+3.35	+2.96	+1.54	+3.54
OPT-1.3b	Vanilla	**56.24**	**66.78**	65.68	**65.42**	73.22	67.84	70.39	69.82	75.38	70.80
	TF-IDF	32.92	41.26	47.99	45.42	56.90	45.73	46.37	45.56	46.82	53.98
	Keywords	31.29	37.59	52.82	53.22	60.67	59.30	60.34	59.76	63.84	64.60
	Summary	27.40	40.03	46.65	50.17	51.46	53.77	54.75	50.89	61.54	55.75
	SelCon	27.61	31.64	37.53	37.29	42.26	38.69	43.58	49.11	46.15	38.05
	LLMLingua	54.81	66.26	66.49	57.29	68.20	64.82	73.18	71.01	63.08	69.91
	Ours	53.41	62.76	**69.71**	63.73	**76.15**	**70.85**	**74.86**	**75.15**	**76.15**	**73.45**
	$\alpha = 0.01$	21.47	35.66	49.06	51.86	50.63	56.78	63.69	62.72	63.85	61.95
	$\alpha = 0.05$	30.06	42.66	54.42	60.00	61.09	60.80	67.04	65.09	65.38	64.60
	$\alpha = 0.1$	34.97	51.57	61.13	59.66	64.44	64.32	68.16	67.46	68.46	64.60
	$\alpha = 0.5$	46.63	58.74	65.15	65.08	64.85	66.33	67.04	68.64	68.46	68.14
	Δ	-2.83	-4.02	+4.03	-1.69	+2.93	+3.01	+4.47	+5.33	+0.77	+2.65
OPT-2.7b	Vanilla	57.46	70.80	72.12	71.53	**76.99**	**78.39**	**77.65**	**79.29**	**82.31**	79.64
	TF-IDF	34.97	41.96	50.67	51.19	63.60	64.82	63.13	64.50	68.46	62.83
	Keywords	33.95	43.01	52.82	59.32	65.69	62.81	69.83	67.46	70.77	73.45
	Summary	27.20	42.13	48.79	52.20	53.56	49.25	49.16	52.07	52.31	48.67
	SelCon	30.06	35.14	43.70	42.37	45.19	44.22	46.37	48.52	49.23	43.36
	LLMLingua	**57.87**	70.80	**75.34**	74.24	74.90	73.37	69.83	68.64	67.69	80.53
	Ours	56.62	**71.50**	74.80	**76.61**	76.57	**78.39**	71.51	71.01	72.31	**82.30**
	$\alpha = 0.01$	25.15	36.19	47.99	56.27	55.23	60.80	68.16	68.64	72.31	69.91
	$\alpha = 0.05$	34.76	45.45	57.10	66.10	64.02	70.85	69.27	71.60	74.62	72.57
	$\alpha = 0.1$	39.47	54.72	58.71	66.10	69.87	69.35	73.18	71.01	74.62	77.88
	$\alpha = 0.5$	53.37	65.21	69.44	74.24	74.48	71.36	69.27	72.78	72.31	74.34
	Δ	-0.84	+0.70	+2.68	+5.08	-0.42	0.00	-6.14	-8.28	-10.00	+2.66
Bloom-560m	Vanilla	**48.26**	**56.47**	53.62	**53.22**	57.32	60.80	**54.75**	**59.17**	**61.53**	**61.95**
	TF-IDF	26.18	27.27	31.37	28.14	41.00	35.18	38.55	43.79	36.15	39.82
	Keywords	24.74	35.31	41.29	41.69	48.54	46.23	54.19	47.33	42.31	46.90
	Summary	21.68	34.97	43.70	40.00	45.19	50.75	51.40	46.75	46.92	51.33
	SelCon	16.77	19.23	25.20	20.68	25.10	30.15	34.08	36.69	34.62	34.51
	LLMLingua	44.38	54.90	**56.03**	48.47	**59.41**	**64.82**	50.84	57.40	52.31	60.18
	Ours	42.97	52.27	53.35	47.12	59.00	64.32	50.28	56.21	60.77	59.29
	$\alpha = 0.01$	17.59	23.25	30.56	31.19	35.56	37.69	42.46	46.15	37.69	55.75
	$\alpha = 0.05$	23.31	30.42	34.32	36.27	43.10	39.70	52.51	55.62	47.69	56.64
	$\alpha = 0.1$	29.24	36.19	39.14	39.66	47.70	43.22	53.63	55.03	48.46	61.06
	$\alpha = 0.5$	36.40	41.08	44.77	47.80	49.37	48.74	52.21	53.25	50.00	60.18
	Δ	-5.29	-4.20	-0.27	-6.10	+1.68	+3.52	-4.47	-2.96	-0.76	-2.66
Bloom-7b1	Vanilla	**58.28**	**74.65**	**74.26**	**76.61**	79.91	**82.41**	**75.98**	**84.62**	83.08	**89.38**
	TF-IDF	37.63	47.03	53.08	61.36	67.36	63.32	67.04	68.05	72.31	68.14
	Keywords	34.56	50.17	56.57	62.71	69.04	67.34	68.72	73.96	69.23	74.33
	Summary	28.63	40.73	49.33	51.86	55.23	64.82	57.54	65.09	66.15	66.37
	SelCon	27.81	36.36	42.36	43.05	46.44	46.73	43.58	53.85	51.54	46.90
	LLMLingua	52.15	71.85	71.58	70.84	82.01	80.40	72.63	81.66	75.38	76.99
	Ours	51.12	71.50	71.31	72.88	**83.26**	81.91	73.74	82.25	**83.84**	84.96
	$\alpha = 0.01$	23.72	34.44	46.38	49.49	48.12	55.78	62.57	66.27	67.69	71.68
	$\alpha = 0.05$	31.29	44.76	54.96	58.64	60.25	63.82	70.95	76.92	71.54	74.34
	$\alpha = 0.1$	36.81	53.67	60.05	65.42	67.78	72.36	72.07	78.70	75.38	75.22
	$\alpha = 0.5$	50.31	60.31	62.47	70.51	74.48	77.39	74.86	79.88	72.31	72.57
	Δ	-7.16	-3.15	-2.95	-3.73	+3.35	-0.50	-2.24	-2.37	+0.76	-4.42
Llama-2-chat-13b	Vanilla	54.40	71.69	71.05	73.22	79.08	76.38	74.30	72.78	71.54	79.64
	TF-IDF	49.28	55.24	69.17	70.51	84.10	78.39	80.45	81.66	80.77	82.30
	Keywords	39.47	53.85	60.86	64.41	67.36	74.37	73.18	80.47	79.23	81.42
	Summary	31.29	43.71	45.30	44.07	50.63	51.76	41.90	52.66	56.15	66.37
	SelCon	37.63	43.18	53.35	56.27	59.41	65.33	64.25	68.05	69.23	68.14
	LLMLingua	59.30	73.25	72.39	80.00	81.59	78.39	79.33	80.47	81.54	84.07
	Ours	64.83	**76.40**	**79.36**	80.34	**84.52**	**84.42**	**85.47**	**86.39**	**86.15**	**86.72**
	$\alpha = 0.01$	36.40	41.61	57.37	61.36	68.20	74.37	73.18	80.47	75.38	72.57
	$\alpha = 0.05$	45.40	55.77	68.90	71.53	80.33	77.39	81.56	83.43	79.23	80.53
	$\alpha = 0.1$	49.69	65.03	74.26	76.61	81.59	82.41	81.56	**86.39**	79.23	81.42
	$\alpha = 0.5$	**65.64**	**76.40**	77.75	**82.03**	84.10	82.91	81.01	81.07	81.54	81.42
	Δ	+10.43	+4.71	+8.31	+7.12	+5.44	+8.04	+11.17	+13.61	+14.61	+7.08
Llama-3.1-8B-Instruct	Vanilla	66.67	81.29	84.18	82.03	82.85	83.42	86.03	85.21	85.38	80.53
	TF-IDF	56.65	61.19	72.39	75.93	78.66	69.85	72.07	69.82	70.00	58.41
	Keywords	55.62	63.81	71.31	71.53	76.15	80.40	86.03	81.66	76.15	78.76
	Summary	39.26	48.08	54.42	55.59	60.25	61.81	55.87	60.36	60.77	69.03
	SelCon	41.41	50.35	61.39	59.66	59.41	68.34	63.12	62.72	68.46	61.95
	LLMLingua	74.44	85.31	86.60	93.56	87.45	89.95	88.83	90.53	89.23	88.50
	Ours	76.89	84.44	91.15	89.49	92.05	93.97	93.30	92.90	90.77	94.69
	$\alpha = 0.01$	45.40	59.62	69.44	74.58	80.75	80.90	87.71	84.02	90.00	86.73
	$\alpha = 0.05$	55.21	69.93	78.55	80.34	87.45	87.44	90.50	88.17	88.46	90.27
	$\alpha = 0.1$	59.71	75.70	83.11	84.07	87.45	86.93	92.74	91.72	88.46	92.92
	$\alpha = 0.5$	76.48	83.91	88.74	90.51	90.79	91.96	89.39	89.94	**90.77**	89.38
	Δ	+10.22	+3.15	+6.97	+7.46	+9.20	+10.55	+7.27	+7.69	+5.39	+14.16
DeepSeek-V2-Lite	Vanilla	19.63	41.78	41.02	61.36	66.95	76.88	**74.30**	**75.73**	**84.62**	80.53
	TF-IDF	6.13	8.74	12.06	16.95	22.18	20.60	20.11	24.85	35.38	33.63
	Keywords	4.29	13.11	20.91	28.47	29.29	29.15	38.55	48.52	49.23	49.56
	Summary	9.82	33.39	41.55	42.37	49.79	56.78	56.98	63.91	63.08	70.80
	SelCon	15.33	31.11	37.80	38.64	48.12	47.74	46.37	55.03	55.38	54.87
	LLMLingua	32.72	**60.31**	**67.56**	**76.95**	79.08	**82.41**	68.72	73.96	83.85	78.76
	Ours	44.38	59.62	66.75	74.58	**80.75**	**82.41**	67.60	69.82	**84.62**	**87.61**
	$\alpha = 0.01$	27.81	39.69	46.92	48.47	56.90	63.82	64.25	59.17	60.77	72.57
	$\alpha = 0.05$	34.56	46.50	53.35	56.27	51.46	59.80	57.54	60.36	60.77	69.03
	$\alpha = 0.1$	37.01	51.92	56.30	56.27	57.74	61.81	65.36	63.91	66.15	73.45
	$\alpha = 0.5$	**46.26**	59.09	62.73	58.64	58.58	74.87	69.83	68.64	77.69	78.76
	Δ	+24.75	+17.84	+25.73	+13.22	+13.80	+5.53	-6.70	-5.91	0.00	+7.08
Qwen3-32B	Vanilla	29.65	28.15	34.05	28.47	25.94	32.66	18.99	17.75	23.08	23.01
	TF-IDF	14.52	18.36	22.79	28.14	31.38	30.15	24.58	29.59	29.23	28.32
	Keywords	25.15	27.79	31.90	32.88	36.40	33.17	30.17	34.32	31.54	32.74
	Summary	25.36	28.32	24.66	28.47	28.03	20.60	20.11	18.34	16.92	23.89
	SelCon	17.79	18.53	26.54	28.14	28.87	23.12	24.58	27.22	30.00	21.24
	LLMLingua	**42.74**	**44.06**	**47.45**	43.39	46.44	39.20	34.64	40.24	35.38	45.13
	Ours	37.22	41.08	43.17	**48.81**	**50.21**	**45.23**	**48.60**	**44.97**	**42.31**	46.02
	$\alpha = 0.01$	20.45	26.92	32.17	36.95	43.93	40.70	41.34	31.36	33.85	43.36
	$\alpha = 0.05$	23.31	33.39	39.95	38.31	42.26	41.71	42.46	42.60	37.69	36.28
	$\alpha = 0.1$	26.58	40.21	39.95	42.03	46.86	42.21	44.13	40.24	37.69	36.28
	$\alpha = 0.5$	36.81	40.03	43.70	45.23	48.12	43.72	34.64	27.81	34.62	38.94
	Δ	+7.57	+12.93	+9.12	+20.34	+24.27	+12.57	+29.61	+27.22	+19.23	+23.01

Table A3: The ablation study results of AUC ↑. The LLMs' order and symbol definitions are the same as Table 2.

Datasets	α	K	G-1.3	G-2.7	O-1.3	O-2.7	b-560	b-7b1	L-13	L3.1-8	DS-V2	Q3-32
PopQA	Ours	I_s	**600.62**	**611.43**	**625.14**	**648.91**	**587.98**	**677.77**	**678.51**	**756.44**	**648.90**	**356.55**
		I_l	**283.54**	**296.09**	298.73	**308.92**	**292.74**	**332.16**	**326.67**	**357.74**	**318.06**	**191.09**
	0.01	I_s	474.99	476.62	513.82	521.05	427.70	521.88	534.01	646.48	430.18	275.57
		I_l	261.01	255.40	286.18	279.83	249.96	285.88	288.66	339.13	231.95	151.76
		Δ_{I_s}	-125.63	-134.81	-111.32	-127.86	-160.28	-155.89	-144.50	-109.96	-218.72	-80.98
		Δ_{I_l}	-22.53	-40.69	-12.55	-29.09	-42.78	-46.28	-38.01	-18.61	-86.11	-39.33
	0.05	I_s	518.36	527.80	555.57	585.21	486.72	593.21	575.42	692.99	444.91	328.01
		I_l	268.39	268.61	290.00	302.73	263.38	306.92	296.35	344.75	228.82	190.62
		Δ_{I_s}	-82.26	-83.63	-69.57	-63.70	-101.26	-84.56	-103.09	-63.45	-203.99	-28.54
		Δ_{I_l}	-15.15	-27.48	-8.73	-6.19	-29.36	-25.24	-30.32	-12.99	-89.24	-0.47
	0.1	I_s	518.36	555.59	590.69	604.98	508.20	611.86	615.68	721.41	484.82	330.48
		I_l	268.39	288.02	**302.36**	304.22	277.27	316.30	305.38	351.58	249.50	182.36
		Δ_{I_s}	-82.26	-55.84	-34.45	-43.93	-79.78	-65.91	-62.83	-35.03	-164.08	-26.07
		Δ_{I_l}	-15.15	-8.07	+3.63	-4.70	-15.47	-15.86	-21.29	-6.16	-68.56	-8.73
	0.5	I_s	550.01	580.26	599.36	619.43	534.50	648.25	658.08	740.35	560.19	300.90
		I_l	269.94	283.66	300.58	299.82	271.24	323.96	315.40	347.72	296.09	147.51
		Δ_{I_s}	-50.61	-31.17	-25.78	-29.48	-53.48	-29.52	-20.43	-16.09	-88.71	-55.65
		Δ_{I_l}	-13.60	-12.43	+1.85	-9.10	-21.50	-8.20	-11.27	-10.02	-21.97	-43.58
EntityQuestions	Ours	I_s	**546.46**	**627.41**	**632.79**	**662.16**	**494.45**	**688.73**	**738.82**	**813.86**	**652.14**	**406.00**
		I_l	**248.82**	**294.48**	**298.31**	**295.18**	**229.06**	**323.26**	**343.58**	**371.30**	**307.05**	**181.50**
	0.01	I_s	398.68	468.53	475.96	513.12	321.22	478.44	586.43	693.08	490.18	319.13
		I_l	213.20	252.50	249.62	274.46	173.02	260.26	302.50	345.54	252.38	148.58
		Δ_{I_s}	-147.78	-158.88	-156.83	-149.04	-173.23	-210.29	-152.39	-120.78	-161.96	-86.87
		Δ_{I_l}	-35.62	-41.98	-48.69	-20.72	-56.04	-63.00	-41.08	-25.76	-54.67	-32.92
	0.05	I_s	461.64	539.16	523.81	572.68	379.61	554.66	661.10	743.58	497.84	348.16
		I_l	235.35	271.86	260.21	287.21	203.99	288.49	323.18	355.98	243.08	161.74
		Δ_{I_s}	-84.82	-88.25	-108.98	-89.48	-114.84	-134.07	-77.72	-70.28	-154.30	-57.84
		Δ_{I_l}	-13.47	-22.62	-38.10	-7.98	-25.07	-34.77	-20.40	-15.32	-63.97	-19.76
	0.1	I_s	501.54	564.93	554.98	596.23	408.18	601.44	692.64	766.50	534.69	364.75
		I_l	248.16	276.75	268.54	292.42	209.26	299.94	329.10	362.84	263.05	161.30
		Δ_{I_s}	-44.92	-62.48	-77.81	-65.93	-86.27	-87.29	-46.18	-47.36	-117.45	-41.25
		Δ_{I_l}	-0.66	-17.73	-29.77	-2.76	-19.80	-23.32	-14.48	-8.46	-44.00	-20.20
	0.5	I_s	491.76	613.74	581.67	632.94	435.51	633.65	720.34	798.94	592.58	355.74
		I_l	226.26	288.91	271.38	287.21	209.92	302.03	325.79	360.77	292.97	138.40
		Δ_{I_s}	-54.70	-13.67	-51.12	-29.22	-58.94	-55.08	-18.48	-14.92	-59.56	-50.26
		Δ_{I_l}	-22.56	-5.57	-26.93	-7.97	-19.14	-21.23	-17.79	-10.53	-14.08	-43.10