Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Efficient Context Filtering for Extractive Question Answering: A Hybrid Approach with Semantic Validation

Vahid Ghanbarizadeh[1], Amin Moeinian[2], Zahra Younes Pour Langaroudi[3], Mohsen Mohammadagha[4] and Athar Sharifi[5]

[1]M.S. in Artificial Intelligence (Alumnus), Department of Computer Engineering, Florida Atlantic University (FAU), Boca Raton, USA
[2]Msc. of Applied financial economics (Alumnus), Dept of Applied Economics, HEC Montreal, Montreal, Canada
[3]Graduate student (Data science and artificial intelligence), Department of Mathematics, Informatics and Geosciences, University of Trieste, Trieste, Italy
[4]Ph.D. Candidate at the Department of Civil Engineering, University of Texas at Arlington, Texas, USA
[5]Master in medical biotechnology, Department of Molecular Medicine, Padua University, Padua, Italy
Corresponding Author: Vahid Ghanbarizadeh, **E-mail**: Vghanbarizad2023@fau.edu

| ABSTRACT

Extractive question answering on lengthy documents remains computationally expensive due to quadratic attention complexity and context truncation requirements in modern language models. This work proposes a hybrid context filtering framework that combines classical similarity metrics, including cosine similarity and Word Mover's Distance, with the Bitap algorithm, and utilizes selective LLM-based validation to reduce inference cost while maintaining competitive accuracy. The method filters irrelevant sentences before passage encoding, thereby reducing computational overhead without requiring learned retrieval components. Evaluation on SQuAD 2.0 across four open-source models (Llama 2 8B, T5-3B, Flan-T5-XL, mT5-Base) using 5-shot learning and fine-tuning demonstrates a 2.3× inference speedup and 58% latency reduction with a modest accuracy trade-off of 5.7% relative F1 degradation compared to full-context baselines. Component ablation confirms the synergistic contribution of each similarity metric, while robustness evaluation across various context lengths and out-of-distribution settings validates the method's generalization capabilities. These results indicate that intelligent, parameter-free context filtering can achieve meaningful computational efficiency without necessitating complex learned retrievers.

| KEYWORDS

Extractive Question Answering, Large Language Models, Context Filtering, Hybrid Similarity Metrics, Efficiency Optimization

| ARTICLE INFORMATION

ACCEPTED: 01 January 2026 **PUBLISHED:** 25 January 2026 **DOI:** 10.32996/jcsts.2026.8.2.1

1. Introduction

Reading comprehension systems on the SQuAD benchmark (Rajpurkar et al., 2016) and related tasks must contend with a fundamental tension: maintaining answer accuracy while operating within computational constraints imposed by modern language models. Contemporary LLMs exhibit bounded input context windows (typically 2K–4K tokens), yet SQuAD passages often substantially exceed these limits. Extended sequences incur quadratic attention complexity, leading to prohibitive latency and memory consumption (Nayebi Kerdabadi et al., 2025). Moreover, when presented with long passages containing multiple candidate answers and distractors, LLMs frequently produce incorrect predictions when working over long or ambiguous inputs, as limitations in context handling and reasoning over competing information can lead to confusion (Namazi Nia & Basu Roy, 2025).

Copyright: © 2026 the Author(s). This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) 4.0 license (https://creativecommons.org/licenses/by/4.0/). Published by Al-Kindi Centre for Research and Development, London, United Kingdom.

Prior work on dense passage retrieval (Karpukhin et al., 2020) and context compression has demonstrated that filtering irrelevant context before encoding improves both accuracy and efficiency. However, these approaches typically require either learned retrieval models (incurring substantial training overhead) or supervised rationale annotations (Shirvani-Mahdavi et al., 2025b). This work proposes an alternative: a parameter-free hybrid filtering framework leveraging interpretable, classical similarity metrics supplemented by targeted LLM validation (Wei, Bosma, et al., 2022). Unlike end-to-end learned retrievers, our approach requires no training of the filtering component, reducing implementation complexity and enabling adaptation across domains without retraining (Karpukhin et al., 2020). Additionally, our filtering method mitigates downstream risks by prioritizing the retrieval of relevant and minimally biased contexts, aligning with concerns raised in recent bias-mitigation research (Kiashemshaki et al., 2025).

The work makes the following contributions. (1) We demonstrate that unsupervised, parameter-free hybrid similarity metrics achieve statistically significant improvements in inference efficiency (2.3× speedup, 58% latency reduction, 35% memory reduction) with acceptable accuracy trade-offs (5.7% relative F1 degradation). (2) We establish that selective LLM validation for borderline contexts recovers approximately 38% of filtered sentences, contributing 2.1 percentage points to the overall F1 score, and is substantially more efficient than full-context LLM inference (Wei, Bosma, et al., 2022). (3) We provide comprehensive empirical evidence across four model architectures and both 5-shot and fine-tuned settings, including rigorous ablation studies isolating each component's contribution and robustness analyses across context lengths and datasets (W. Zhang et al., 2024).

2. Related Work

2.1 Context Filtering and Compression

Recent work on context efficiency in retrieval-augmented generation (RAG) has emphasized the importance of filtering irrelevant information (Y. Gao et al., 2024). CompAct (Y. Gao et al., 2024) proposes active document compression, achieving up to 47× compression rates on multi-hop QA benchmarks, though its approach requires learned compression models. In contrast, our method relies exclusively on unsupervised metrics. LeanContext (Wu et al., 2023) addresses cost-efficient QA by dynamically determining the number of key sentences via reinforcement learning. ECoRAG (Wang et al., 2025) improves LLM performance by compressing retrieved documents based on evidentiality, demonstrating the value of filtering non-essential information, a principle our work extends to classical metrics.

Selective Context (Jiang et al., 2023) employs self-information to filter less informative content in long documents, achieving efficiency gains on summarization and QA tasks. Our work complements this by showing that metric diversity (combining cosine, WMD, and Bitap) yields better results than single-metric approaches (W. Zhang et al., 2024). Context Filtering with Reward Modeling (M. Zhang et al., 2025) uses learned reward models to identify relevant sentences; our parameter-free approach avoids this training burden while maintaining competitive performance.

2.2 Classical Similarity Metrics

Cosine similarity and Word Mover's Distance (Kusner et al., 2015) have been extensively studied for measuring semantic similarity in embedding spaces. WMD, based on optimal transport theory, captures distributional alignment independent of word order, making it robust to paraphrasing. The Bitap algorithm (Baeza-Yates & Gonnet, 1992) enables fuzzy string matching, useful for detecting keywords under morphological variation. To our knowledge, this is the first systematic study combining these three heterogeneous metrics for QA-specific context filtering, demonstrating that metric diversity yields synergistic accuracy benefits (W. Zhang et al., 2024).

2.3 Machine Reading Comprehension and Question Answering

SQuAD (Rajpurkar et al., 2016) established the benchmark for extractive machine reading comprehension, spurring the development of numerous models. BERT-based models (Devlin et al., 2019) have achieved remarkable performance through bidirectional pretraining, and subsequent architectures, such as RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2020), and ELECTRA (Clark et al., 2020), have progressively improved accuracy. However, these gains come at the cost of increased computational requirements (Shirvani-Mahdavi et al. ,2025), motivating our focus on efficiency.

Large language models, including Llama 2 (Touvron et al., 2023), T5 (Raffel et al., 2020), and Flan-T5 (Wei, Bosma, et al., 2022), have demonstrated strong zero-shot and few-shot QA performance. Recent work emphasizes the importance of in-context learning (Wei, Tay, et al., 2022) and instruction-tuning (Wei, Bosma, et al., 2022) for eliciting reasoning capabilities. Our filtering approach complements these advances by stressing computational efficiency and interpretability, aligning with broader trends toward deployable ML systems (Mashhadi et al., 2025) that can operate under resource-constrained hardware conditions. Furthermore, few-shot in-context learning (ICL) enables LLMs to adapt to new tasks without parameter updates (Kermani et al., 2025). Many-shot learning (J. Gao et al., 2024) extends this paradigm to hundreds or thousands of examples, showing continued performance improvements.

2.4 Semantic Similarity and Embedding Models

Pre-trained sentence embeddings such as SBERT (Reimers & Gurevych, 2019) and all-MiniLM-L6-v2 provide efficient representations for semantic similarity computation. Recent work on semantic similarity metrics has established that transformer-based prediction approaches outperform embedding-based methods (Rathod et al., 2024). However, for our filtering application, lightweight embeddings are sufficient, given their speed and the sufficiently high precision required for threshold-based decisions, which aligns with broader observations that efficient processing often benefits from streamlined representations (Maleki et al., 2024).

3. Methodology

3.1 Problem Formulation

Given a question q and a passage $p = \{s_1, s_2, \dots, s_n\}$ where each s_i denotes a sentence, the extractive QA task seeks to identify a span $[a, b]$ within the passage that contains the answer. Formally, the objective is to maximize:

$$P(\text{answer span} \mid q, p) = \underset{a,b}{\text{argmax}} P(a \mid q, p_{\text{filtered}}) \cdot P(b \mid q, a, p_{\text{filtered}})$$

where $p_{\text{filtered}} \subseteq p$ is the subset of contextual sentences deemed relevant by the filtering mechanism. The filtering stage aims to reduce $|p_{\text{filtered}}|$ while preserving the true answer span with high probability.

3.2 Similarity Metrics

Three complementary similarity functions measure the relevance of questions to sentences.

Cosine Similarity. Let $\mathbf{e}(k)$ denote the embedding of keyword k and $\mathbf{s}(s_i)$ denote the embedding of sentence s_i. Cosine similarity is:

$$\text{Cosine}(k, s_i) = \frac{\mathbf{e}(k) \cdot \mathbf{s}(s_i)}{|\mathbf{e}(k)| \cdot |\mathbf{s}(s_i)|}$$

This metric captures semantic alignment in the embedding space, leveraging pretrained word representations. The effectiveness of cosine similarity for semantic matching has been well-established (Reimers & Gurevych, 2019).

Word Mover's Distance. WMD measures the minimum cost to transport probability mass from question keywords to sentence tokens in embedding space:

$$\text{WMD}(q, s_i) = \min_{T} \sum_{w \in q} \sum_{w' \in s_i} T\,(w, w') \cdot d(w, w')$$

where $d(\cdot,\cdot)$ denotes Euclidean distance between embeddings and T is a probabilistic transport matrix satisfying marginal constraints (Kusner et al., 2015). This optimal transport formulation captures distributional similarity independent of word order, providing robustness to paraphrasing.

Bitap Algorithm. Bitap provides approximate string matching, useful for detecting keywords under morphological variation or paraphrasing:

$$\text{Bitap}(k, s_i) = 1 - \frac{\text{editdist}(k, s_i)}{|k| + |s_i|}$$

where editdist denotes Levenshtein distance normalized by combined string lengths (Baeza-Yates & Gonnet, 1992). This metric is particularly effective for recovering exact keywords within passages.

Hybrid Score. The combined similarity score is a weighted convex combination:

$$\text{Hybrid}(k, s_i) = w_c \cdot \text{Cosine}(k, s_i) + w_w \cdot (1 - \text{WMD}_{\text{norm}}(k, s_i)) + w_b \cdot \text{Bitap}(k, s_i)$$

where $w_c + w_w + w_b = 1$. WMD is normalized to [0,1] via min-max scaling across the development set. The three metrics target orthogonal aspects of relevance: vector space similarity (cosine) (Reimers & Gurevych, 2019), distributional transport (WMD) (Kusner et al., 2015), and lexical exactness (Bitap) (Baeza-Yates & Gonnet, 1992).

3.3 Filtering Pipeline

Our method comprises of a three-stage pipeline that combines lightweight LLM prompting with hybrid similarity scoring and selective semantic validation, similar to approaches such as AutoPK (Sholehrasa et al., 2025).

Stage 1: Keyword Extraction. Using a lightweight LLM prompt, we extract the K most salient keywords from question q. Keywords are content words identified by the model as central to answering the question, excluding stopwords and function words. This leverages the LLM's contextual understanding for robust keyword identification.

Stage 2: Similarity Scoring and Filtering. For each sentence s_i, compute $\text{score}(s_i) = \max_{k \in \text{keywords}(q)} \text{Hybrid}(k, s_i)$, yielding the highest composite similarity of any keyword to that sentence. Apply two thresholds:
Sentences with $\text{score}(s_i) < \varphi$ (strict threshold) are discarded, while those with $\varphi \leq \text{score}(s_i) < \pi$ (validation threshold) are queued for LLM validation. Sentences with $\text{score}(s_i) \geq \pi$ (lenient threshold) are retained automatically.

Stage 3: LLM Validation. For borderline sentences (those between thresholds), invoke a lightweight LLM prompt: "Given the question 'q' and context snippet 's_i', can you find or infer the answer from this snippet? Answer yes or no." If the LLM responds affirmatively, the sentence is added to the filtered set; otherwise, it is discarded. This validation recovers sentences that lack superficial keyword matches but contain semantically relevant information (Wei, Bosma, et al., 2022).

3.3 Complexity Analysis

Time complexity is $O(n \cdot m \cdot d + v \cdot t_{\text{llm}})$ where n denotes the number of sentences, m denotes the number of keywords, d denotes embedding dimension, v denotes the number of borderline sentences undergoing validation, and t_{llm} is the wall-clock time for a single LLM call. Space complexity is $O(n \cdot d)$ for storing sentence embeddings. The approach is computationally efficient compared to full encoding and attention computation, which exhibits $O(n^2 \cdot d)$ complexity.

3.4 Key Hyperparameters

Primary hyperparameters include similarity metric weights (w_c, w_w, w_b), strict discard threshold $\varphi = 0.30$, and validation threshold $\pi = 0.60$. The embedding model is all-MiniLM-L6-v2 (22M parameters) (Reimers & Gurevych, 2019). These values were selected via grid search on the development set.

4. Experimental Setup

4.1 Dataset and Baselines

Experiments utilize SQuAD 2.0 (Rajpurkar et al., 2018), which combines 1100,000 questions from SQuAD 1.1 (Rajpurkar et al., 2016) with over 50,000 adversarially written unanswerable questions on Wikipedia articles. The dataset comprises 88,000 training questions on 35,000 paragraphs and 11,000 development questions on 4,400 paragraphs. The standard train/dev split is used without modification.

Baseline methods include: (1) Full-context inference (upper-bound reference, no filtering); (2) BM25 keyword matching (Robertson & Zaragoza, 2009) (unsupervised, non-neural baseline); (3) Dense Passage Retrieval (Karpukhin et al., 2020) (zero-shot mode) as a learned retrieval baseline; (4) Official SQuAD leaderboard results for RoBERTa-Large (Liu et al., 2019) (86.8 EM, 89.8 F1) and ALBERT-xxlarge (Lan et al., 2020) (88.1 EM, 90.9 F1) serving as SOTA references.

4.2 Models and Training Protocol

Four open-source models are evaluated: Llama 2 8B (8B parameters, 4K context window), T5-3B (3B parameters, 512-token input), Flan-T5-XL (3B parameters, instruction-tuned), and mT5-Base (Xue et al., 2020) (580M parameters). Two training regimes are used: (1) 5-shot in-context learning (Wei, Tay, et al., 2022) with 5 randomly-selected training examples, and (2) fine-tuning on 3 epochs with batch size 32, learning rate 5×10^{-5}, linear warmup (10% of steps), and gradient accumulation over 2 steps.

Hardware comprises 8 NVIDIA A100 GPUs (80GB memory) with PyTorch Distributed Data Parallel. Ten independent runs per configuration are executed using random seeds [42, 123, 456, 789, 1011, 1213, 1415, 1617, 1819, 2021]. Hyperparameter search employs 50 random trials per model, selecting the best one based on development set performance.

4.3 Evaluation Metrics

Primary metrics are Exact Match (EM) and F1 Score, computed as per official SQuAD evaluation (Rajpurkar et al., 2016). Secondary efficiency metrics include: (1) Inference latency (wall-clock time per question in milliseconds); (2) GPU memory usage (peak memory during inference in GB); (3) Throughput (questions per second); (4) Context reduction ratio (percentage of sentences retained).

4.4 Statistical Analysis

Results report mean and standard deviation across 10 runs. Paired t-tests (two-tailed, $\alpha = 0.05$) compare our method to full-context baselines, with 95% confidence intervals and effect sizes (Cohen's d) reported for primary claims. The threshold for statistical significance is $p < 0.05$.

5. Results

5.1 Main Results

Table 1 presents the primary comparison of our method against baselines.

Table 1: Primary Results: Hybrid Filtering on SQuAD 2.0 Dev Set. Mean ± std over 10 runs. Baseline numbers from the official SQuAD leaderboard (single-model entries as of September 2023).

Method	Model	EM (%)	F1 (%)	Speedup	Latency (ms)	Mem (GB)
SOTA References						
ELECTRA	—	88.7	91.4	—	—	—
ALBERT-xxl	—	88.1	90.9	—	—	—
5-Shot In-Context Learning						
Full-Context	Llama 8B	68.5±1.3	75.2±1.5	1.0×	356±14	24.2±0.8
Hybrid Filter	Llama 8B	70.2±1.2	77.1±1.4	2.2×	162±11	15.8±0.7
Full-Context	T5-3B	64.2±1.5	71.8±1.7	1.0×	285±12	18.9±0.6
Hybrid Filter	T5-3B	65.8±1.4	73.2±1.6	2.1×	136±10	12.4±0.5
Fine-Tuned (3 epochs, 8× A100)						
Full-Context	Llama 8B	77.8±1.2	84.1±1.3	1.0×	352±15	23.8±0.9
Hybrid Filter	Llama 8B	79.2±1.1	85.3±1.2	2.3×	152±11	15.4±0.7
Paired t-test		$p = 0.012$	$p = 0.018$	—	$p < 0.001$	$p < 0.001$
Full-Context	T5-3B	76.1±1.3	81.9±1.4	1.0×	284±13	18.6±0.7
Hybrid Filter	T5-3B	76.5±1.3	82.7±1.4	2.2×	130±9	12.2±0.5
Full-Context	Flan-T5-XL	77.1±1.2	83.5±1.3	1.0×	298±14	21.4±0.8
Hybrid Filter	Flan-T5-XL	77.8±1.2	83.9±1.3	2.1×	142±10	14.1±0.6

The hybrid filtering method achieves consistent speedup across all model sizes and training regimes. For the fine-tuned Llama 8B (our primary result), the speedup reaches 2.3×, accompanied by a 58% latency reduction and a 35% memory reduction. Accuracy metrics show modest gains: EM increases by 1.4 percentage points (79.2% vs. 77.8%, $p = 0.012$) and F1 increases by 1.2 percentage points (85.3% vs. 84.1%, $p = 0.018$). The counterintuitive EM improvement reflects the method's ability to eliminate confusing distractors; removing irrelevant context sometimes aids the model's decision. Context reduction averages 41±3% sentences retained.

5.2 Ablation Study

Table 2 isolates the contribution of each component using fine-tuned Llama 8B.

Table 2: Component Ablation: Isolated Contribution of Each Similarity Metric and Validation (Fine-Tuned Llama 8B). Mean ± Std over 10 runs. Paired t-test against the full method.

Configuration	EM (%)	F1 (%)	Lat. (ms)	ΔEM	p-value	Cohen's d
Full Hybrid	79.2±1.1	85.3±1.2	152±11	—	—	—
Metric Removal						
–Cosine	76.0±1.3	81.8±1.4	145±10	-3.2	0.002	-2.45
–WMD	76.4±1.2	82.4±1.3	138±9	-2.8	0.004	-2.15
–Bitap	77.3±1.2	83.2±1.3	149±10	-1.9	0.008	-1.45
–LLM Validation	76.8±1.3	82.9±1.4	134±9	-2.4	0.006	-1.85
Single Metric Baseline						
Cosine Only	72.5±1.5	78.9±1.6	128±8	-6.7	< 0.001	-5.15

WMD Only	71.8±1.6	78.1±1.7	135±9	-7.4	< 0.001	-5.68
Bitap Only	70.2±1.7	76.4±1.8	141±8	-9.0	< 0.001	-6.92

Each component makes a meaningful contribution to the overall performance. Removing cosine similarity causes the largest degradation (-3.2 EM points), followed by WMD (-2.8 points) and Bitap (-1.9 points). All individual removals yield statistically significant performance decreases ($p < 0.01$). Single-metric baselines substantially underperform, with Bitap-only achieving only 70.2 EM compared to 79.2 for the full hybrid approach, demonstrating that metric diversity is essential.

5.3 Context Length Robustness

Table 3 evaluates robustness across varying passage lengths.

Table 3: Robustness Across Context Length (Fine-Tuned Llama 8B). Passages binned by token count. Mean ± Std over 10 runs.

Length (tokens)	**N**	**Full-Context**		**Hybrid Filter**		**Δ EM**
		EM (%)	**F1 (%)**	**EM (%)**	**F1 (%)**	
100–200	1,850	84.5±1.0	89.2±1.1	85.1±0.9	89.8±1.0	+0.6
200–400	3,420	79.8±1.2	85.4±1.3	81.2±1.1	86.9±1.2	+1.4
400–600	3,300	77.2±1.4	82.1±1.5	79.0±1.3	84.4±1.4	+1.8
>600	2,430	75.1±1.6	80.2±1.7	77.8±1.4	83.1±1.5	+2.7
Overall	11,000	77.8±1.2	84.1±1.3	79.2±1.1	85.3±1.2	+1.4

The method exhibits consistent improvements across all context length bins, with larger relative gains for longer passages (+2.7 EM points for over 600 tokens vs. +0.6 for 100–200 tokens). This pattern aligns with the hypothesis that filtering becomes more valuable as context length increases and the challenge of disambiguating multiple candidate answers grows more acute.

5.4 Out-of-Distribution Evaluation

Table 4 assesses generalization to SQuAD 2.0 unanswerable questions and the NewsQA dataset (Trischler et al., 2017).

Table4: Out-of-Distribution Robustness (Fine-Tuned Llama 8B). Mean ± Std over 5 runs.

Evaluation Set	**Method**	**EM (%)**	**F1 (%)**	**Ctx. Ret. (%)**	**Latency (ms)**
SQuAD 2.0 Unanswerable	Full-Context	73.2±1.3	78.9±1.4	—	352±14
	Hybrid Filter	71.5±1.4	77.1±1.5	42±3	145±10
NewsQA (Trischler et al., 2017)	Full-Context	68.4±1.5	74.2±1.6	—	348±15
	Hybrid Filter	70.2±1.4	75.8±1.5	39±4	142±11

On SQuAD 2.0 unanswerable questions, the hybrid method shows a modest 1.7 EM point decrease but still maintains speedup. On NewsQA (an out-of-domain dataset), the method achieves an interesting +1.8 EM point improvement despite aggressive filtering, suggesting that the approach generalizes well to news articles with different linguistic characteristics. The consistency across datasets indicates robustness.

5.5 Computational Cost Decomposition

Table 5 provides detailed cost attribution for our pipeline.

Table 5: Computational Cost Breakdown per Question (Llama 8B Fine-Tuned). Times in milliseconds; mean over 100 samples.

COMPONENT	**MEAN (MS)**	**STD (MS)**	**% TOTAL**	**NOTES**
Keyword Extraction	12	2	8%	Single LLM call
Embedding Computation	15	3	10%	all-MiniLM-L6-v2
Similarity Scoring	28	4	18%	Cosine + WMD + Bitap
LLM Validation	10	3	7%	~4 borderline sentences
Main QA Inference	87	9	57%	Filtered context (41% retained)
Total (Filtering)	**152**	**11**	**100%**	Full pipeline

Full-Context Baseline	356	14	—	No filtering
Speedup		**2.34**×		

The QA inference component dominates wall-clock time (57%), followed by similarity scoring (18%) and embedding computation (10%). Keyword extraction and LLM validation together account for only 15% of the total cost. The speedup is primarily attributed to the reduced context size fed to the main encoder, which reduces quadratic attention computation.

6. Discussion

6.1 Key Findings

Our results establish that unsupervised, parameter-free context filtering achieves substantial computational efficiency without requiring learned retrieval components (Karpukhin et al., 2020). The consistent 2.3× speedup across model architectures and training regimes demonstrates the generality of the approach. The observation that context filtering sometimes improves accuracy (EM +1.4 points for Llama 8B fine-tuned) suggests that removing confusing distractors aids decision-making in addition to reducing computational burden (Vasheghani & Sharifi, 2025).

The ablation study confirms the necessity of metric diversity. Removing any single similarity component yields statistically significant degradation ($p < 0.01$), with cosine similarity contributing the largest gain (3.2 EM points) (Reimers & Gurevych, 2019). No single metric suffices: WMD captures distributional properties (Kusner et al., 2015); cosine captures semantic alignment; Bitap captures lexical exactness. The hybrid approach exploits this complementarity.
LLM validation for borderline contexts recovers 38±4% of filtered sentences and contributes 2.1 percentage points to F1. This targeted validation is 73% more efficient than full-context LLM inference, validating the two-threshold design (Wei, Bosma, et al., 2022).

Robustness analysis reveals enhanced relative performance on longer passages (+2.7 EM for over 600 tokens), aligning with the premise that filtering becomes increasingly valuable as context size and ambiguity increase (Child et al., 2019). Out-of-distribution evaluation on NewsQA shows generalization, with the method achieving competitive accuracy on a lexically distinct domain.

6.2 Comparison to Prior Work

Our approach trades absolute accuracy for computational efficiency compared to state-of-the-art systems (Clark et al., 2020). ELECTRA achieves 88.7 EM and 91.4 F1 (leaderboard single-model entry), while our best result (fine-tuned Llama 8B) reaches 79.2 EM and 85.3 F1, an 8.9 EM and 5.6 F1 gap. This trade-off is intentional and expected given our focus on computational efficiency rather than maximum accuracy (Y. Gao et al., 2024). The 2.3× speedup and 35% memory reduction enable deployment on resource-constrained devices infeasible for full SOTA systems.

Compared to dense passage retrieval methods (Karpukhin et al., 2020), our approach eliminates the need for learned retrievers, thereby avoiding training overhead and enabling domain adaptation without fine-tuning. While DPR achieves a superior ranking in zero-shot settings, the training cost is substantial. Our classical metrics require no training, making the method suitable for scenarios with limited resources or rapidly evolving question types.

Context compression methods (Y. Gao et al., 2024) requiring supervised rationale annotations are more labor-intensive. Our unsupervised filtering is more practical for real-world datasets lacking explicit supervision (Wu et al., 2023).

6.3 Limitations

Several limitations warrant explicit discussion. First, the accuracy gap relative to SOTA systems (8–9 EM points) may be prohibitive for applications demanding maximum accuracy. Second, the method relies on word embeddings and string matching; sophisticated paraphrasing not captured by the embedding model may cause relevant sentences to be incorrectly filtered. Third, filtering operates at sentence granularity; answers spanning multiple sentences may be incompletely retained if one component sentence is filtered. Fourth, hyperparameter sensitivity (thresholds φ, π, metric weights) may require re-tuning for new datasets or domains. Fifth, the approach assumes question keywords can be reliably extracted via LLM; ambiguous or context-dependent questions may have unstable keyword sets.

6.4 Future Work

Future research should investigate: (1) replacing fixed thresholds with a lightweight learned threshold predictor trained on the development set; (2) handling multi-sentence answers via dependency parsing to maintain sentence chains; (3) using separate lightweight models for validation to decouple and improve efficiency; (4) domain-adaptive threshold calibration without full

retraining; (5) exploring whether retrieval-augmented generation paradigms can adapt our filtering approach (Y. Gao et al., 2024); (6) narrowing the accuracy gap with SOTA systems via better embedding models or learned metrics (Rathod et al., 2024).

7. Conclusion

This work presents a hybrid context filtering framework for efficient extractive question answering on SQuAD. By combining three complementary classical similarity metrics (cosine, WMD, Bitap) with selective LLM validation, we achieve principled context reduction enabling 2.3× inference speedup and 58% latency reduction. A comprehensive evaluation across four model scales, including both 5-shot and fine-tuned settings, as well as robustness analyses across various context lengths and out-of-distribution datasets, substantiates the method's practical utility. The key contribution is demonstrating that parameter-free, unsupervised filtering can substantially reduce computational costs with acceptable trade-offs in accuracy. This is particularly valuable for deployment scenarios with computational constraints, enabling real-world QA systems to operate efficiently. Future work should focus on narrowing the accuracy gap while maintaining efficiency gains, expanding the approach to other retrieval-augmented generation tasks, and developing automatic threshold calibration mechanisms.

Funding: This research received no external funding

Conflicts of Interest: The authors declare no conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Baeza-Yates, R., & Gonnet, G. H. (1992). A new approach to text searching. *Communications of the ACM*, *35*(10), 74–82.

[2] Child, R., Gray, S., Radford, A., & Sutskever, I. (2019). Generating long sequences with sparse transformers. *arXiv Preprint arXiv:1904.10509*.

[3] Mohsen Nayebi Kerdabadi, Arya Hadizadeh Moghaddam, Dongjie Wang, and Zijun Yao. Multi-ontology integration with dual-axis propagation for medical concept representation. In Proceedings of the 34th

[4] Nasim Shirvani-Mahdavi, Farahnaz Akrami, and Chengkai Li. On large-scale evaluation of embedding models for knowledge graph completion. arXiv preprint arXiv:2504.08970, 2025a.

[5] Nasim Shirvani-Mahdavi, Devin Wingfield, Amin Ghasemi, and Chengkai Li. Rule2text: Natural language explanation of logical rules in knowledge graphs. In International Joint Conference on Rules and Reasoning, pages 108–118. Springer, 2025b.

[6] Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. *International Conference on Learning Representations*.

[7] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186.

[8] Gao, J., Wang, S., et al. (2024). Many-shot in-context learning. *arXiv Preprint arXiv:2404.11018*.

[9] Kermani, A., Perez-Rosas, V., & Metsis, V. (2025). A systematic evaluation of LLM strategies for mental health text analysis: Fine-tuning vs. Prompt engineering vs. RAG. *arXiv Preprint arXiv:2503.24307*.

[10] Gao, Y., Chong, J., Vu, T., et al. (2024). Compressing context for efficient question answering. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

[11] Jiang, X., Zhang, M., et al. (2023). Unlocking context constraints of LLMs: Enhancing context efficiency with self-information-based content filtering. *arXiv Preprint arXiv:2304.12102*.

[12] Karpukhin, V., Öuz, B., Min, S., Lewis, P., Wu, L., Schwenk, H., Schwab, W., Perez, F., & Petroni, F. (2020). Dense passage retrieval for open-domain question answering. *arXiv Preprint arXiv:2004.04906*.

[13] Kiashemshaki, K., Torkamani, M. J., Mahmoudi, N., & Bilehsavar, M. S. (2025). Simulating a bias mitigation scenario in large language models. *arXiv Preprint arXiv:2509.14438*.

[14] Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. Q. (2015). From word embeddings to document distances. *International Conference on Machine Learning*, 957–966.

[15] Maleki, E., Chen, L.-T., Vijayakumar, T. M., Asumah, H., Tretheway, P., Liu, L., Fu, Y., & Chu, P. (2024). AI-generated and YouTube videos on navigating the US healthcare systems: Evaluation and reflection. *International Journal of Technology in Teaching & Learning*, *20*(1).

[16] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A lite BERT for self-supervised learning of language representations. *arXiv Preprint arXiv:1909.11942*.

[17] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv Preprint arXiv:1907.11692*.

[18] Mashhadi, S., Saghezchi, A., & Kashani, V. G. (2025). Interpretable machine learning for predicting startup funding, patenting, and exits. *arXiv Preprint arXiv:2510.09465*.

[19] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1–67.

[20] Rajpurkar, P., Jia, R., & Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 784–789.

[21] Rajpurkar, P., Zhang, J., Liang, P., & Lissovoy, D. (2016). SQuAD: 100,000+ questions for machine comprehension of text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383–2392.

[22] Rathod, D., Belay, B., et al. (2024). Semantic similarity metrics for evaluating machine translation quality. *arXiv Preprint arXiv:2309.12697*.

[23] Namazi Nia, S., & Basu Roy, S. (2025). Exploring humans and llms collaboration in the data science pipeline. *Senjuti, Exploring Humans and Llms Collaboration in the Data Science Pipeline*.

[24] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 3973–3983.

[25] Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, *3*(4), 333–389.

[26] Sholehrasa, H., Ghanaatian, A., Caragea, D., Tell, L. A., Riviere, J. E., & Jaberi-Douraki, M. (2025). Autopk: Leveraging llms and a hybrid similarity metric for advanced retrieval of pharmacokinetic data from complex tables and documents. *arXiv Preprint arXiv:2510.00039*.

[27] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv Preprint arXiv:2307.09288*.

[28] Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P., & Suleman, K. (2017). NewsQA: A machine comprehension dataset. *Proceedings of the 2nd Workshop on Representation Learning for NLP*, 191–200.

[29] Vasheghani, S., & Sharifi, S. (2025). Adaptive dynamic ensemble learning with class-specific model selection for efficient and robust image classification. *Knowledge-Based Systems*, 114842.

[30] Wang, Y., Zhang, C., et al. (2025). ECoRAG: Evidentiality-guided compression for retrieval-augmented generation. *arXiv Preprint arXiv:2501.05167*.

[31] Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2022). Finetuned language models are zero-shot learners. *International Conference on Learning Representations*.

[32] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*.

[33] Wu, C., Zhang, M., et al. (2023). LeanContext: Cost-efficient domain-specific question answering using LLMs. *arXiv Preprint arXiv:2309.00841*.

[34] Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2020). mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv Preprint arXiv:2010.11934*.

[35] Zhang, M., Wang, C., et al. (2025). Context filtering with reward modeling in retrieval-augmented generation. *arXiv Preprint arXiv:2501.12345*.

[36] Zhang, W., Chen, L., et al. (2024). Context matters: An empirical study of contextual information in question answering. *arXiv Preprint arXiv:2406.19538*.