From Documents to KBLaM: An End-to-End Pipeline for Structured Knowledge Extraction and Injection

Lev Zyurikov, Maria Khodorchenko
ITMO University
Saint-Petersburg, Russia

***Abstract*—Large Language Models demonstrate strong reasoning and language understanding capabilities, yet their ability to incorporate external knowledge remains limited by scalability, efficiency, and robustness, as current approaches require lengthy prompts or costly retrieval at inference time. Knowledge Base augmented Language Models (KBLaM) address these issues by embedding structured knowledge directly into the attention mechanism, enabling linear scaling with respect to knowledge base size and avoiding retrieval-time overhead. However, existing KBLaM evaluations are largely restricted to synthetic or controlled datasets, leaving open the question of their effectiveness in real-world settings.**

In this work, we propose a data-centric pipeline that converts unstructured, real-world documents into KBLaM-compatible knowledge units (name, property, value) while preserving essential semantic content. We evaluate the approach on three realistic knowledge-intensive benchmarks (ObliQA, MultiRC, DROP), analyzing accuracy, coverage, and clarity under varying data conditions. An ablation study confirms that the observed improvements stem from semantically aligned structured knowledge rather than incidental factors. The results demonstrate that the proposed pipeline enables effective adaptation of KBLaM to real-world data, narrowing the gap between theoretical knowledge integration and practical deployment.

I. Introduction

Large Language Models (LLMs) have become a central tool in modern natural language processing, achieving impressive results in question answering, text generation, and reasoning. Despite these advances, LLMs are still limited in their ability to incorporate external knowledge, particularly when such knowledge is large-scale or domain-specific.

Existing solutions typically rely on retrieval-based mechanisms, such as Retrieval-Augmented Generation (RAG), which dynamically fetch relevant information at inference time. While effective, these methods introduce additional system complexity, and their performance degrades as context length increases. Alternatively, in-context learning encodes knowledge directly in the prompt but suffers from poor scalability due to quadratic attention costs.

KBLaM offers a promising alternative by embedding structured knowledge directly into the attention mechanism as continuous key-value representations. This design enables linear scaling with respect to knowledge base size and avoids external retrieval during inference. However, current KBLaM evaluations have been largely confined to synthetic or curated datasets, leaving open the question of their robustness in real-world settings.

While KBLaM assumes access to a clean, structured knowledge base, real-world knowledge is typically embedded in long, heterogeneous documents where facts are implicit, redundant, and inconsistently phrased. Naively converting such data into (name, property, value) triples often introduces noise through spurious entities or attributes, loses the context needed for disambiguation, and does not scale due to the combinatorial growth of candidate facts. This makes KBLaM's practical applicability less about model design and more about reliable, semantics-preserving data preparation.

To address this bottleneck, we present a data-centric pipeline that extracts and decomposes document evidence into minimal, KBLaM-compatible knowledge units. We validate its utility through targeted ablations and benchmark evaluation on three realistic datasets, isolating when structured knowledge genuinely improves question answering under real-data conditions.

Our contributions are the following:

- We propose an end-to-end pipeline that converts unstructured QA-style documents into KBLaM-ready structured knowledge units (name-property-value), enabling attention-based knowledge injection.
- We adapt and evaluate KBLaM on three realistic knowledge-intensive benchmarks (ObliQA, MultiRC, DROP) and report consistent improvements over the base model using both standard QA metrics and an LLM-based quality evaluation (correctness, coverage, clarity).
- We provide diagnostic evidence via ablations (shuffled and empty knowledge) and extraction-quality analysis, showing that gains come from semantically aligned structured knowledge and highlighting current failure modes (e.g., numeric and discrete reasoning).

II. Related Work

A common way to equip LLMs with up-to-date or domain knowledge is to retrieve relevant evidence at inference time. Retrieval-Augmented Generation (RAG) combines parametric generation with non-parametric document retrieval, typically implemented with dense retrievers such as DPR, and can be trained end-to-end as in REALM [1]–[3]. More recent

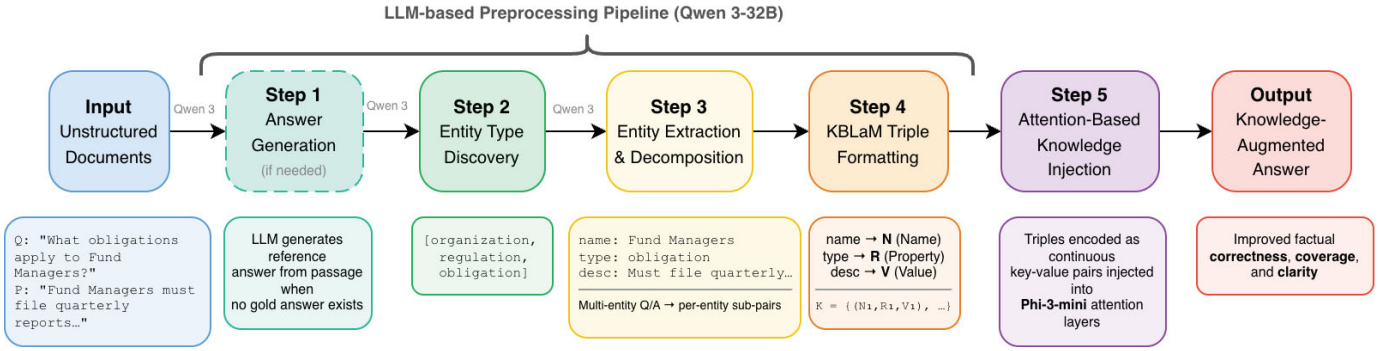


Fig. 1. Overview of the proposed data-centric preprocessing pipeline for adapting real-world documents to KBLaM. Depending on dataset availability, the pipeline first generates reference answers, then discovers relevant entity types, extracts entity-centered information, decomposes multi-entity question-answer pairs into per-entity units, and finally converts the result into KBLaM-compatible $(name, property, value)$ triples used for attention-based knowledge injection.

retrieval-heavy variants (e.g., large-scale nearest-neighbor retrieval over massive text corpora) further improve factual coverage but still introduce retrieval latency and additional system complexity [4]. In contrast, long-context in-context learning can inject evidence by concatenating documents into the prompt, but its compute and memory cost grows quickly with context length, and it remains sensitive to ordering and irrelevant content.

Another direction is to incorporate new knowledge through fine-tuning. Parameter-efficient methods such as LoRA and quantization-aware variants like QLoRA reduce adaptation cost [5], [6], but updating knowledge still requires training and careful data curation, and can risk catastrophic interference; moreover, the resulting knowledge is less interpretable and harder to surgically revise than explicit external stores.

Our work is most closely related to plug-and-play approaches that integrate structured knowledge without retraining the base model. KBLaM proposes to transform a knowledge base into continuous key-value pairs and integrate them via a specialized attention mechanism, enabling dynamic KB updates without running an explicit retriever [7]. AtlasKV extends this line by targeting much larger knowledge graphs under practical GPU constraints [8]. However, these methods largely assume that a reasonably clean, canonicalized KB (often in triple form) already exists. In real deployments, constructing such KBs from noisy documents requires entity normalization, de-duplication, schema alignment, and error control - issues studied broadly in knowledge graph construction and curation [9], [10]. This paper complements KBLaM-style model-side integration by focusing on the data-side pipeline needed to reliably convert real-world corpora into KBLaM-compatible knowledge units and to quantify the impact of common data imperfections.

III. Problem Formalization

Given a set of real-world textual documents, each consisting of a passage and a corresponding question, $D = \{(Q_i, P_i)\}_{i=1}^{N}$, the objective is to transform these data into a representation K compatible with the KBLaM framework. The current KBLaM implementation expects knowledge tokens to be represented as triples of the form (name, property, value), yielding $K = \{(N_j, R_j, V_j)\}_{j=1}^{M}$, where N_j denotes the entity name, R_j the property (relation or attribute), and V_j the corresponding value. The central challenge addressed in this work is to construct a mapping $D \to K$ that converts real-world, unstructured data into this structured format while preserving essential semantic and relational information.

Running example. Consider a regulatory QA instance asking which maturity conditions trigger a disclosure requirement under specific rules. The supporting evidence is typically a short clause embedded in a longer passage, yet KBLaM requires a compact set of knowledge units in the form of triples (N, R, V). For this example, the pipeline extracts two entities (e.g., *Financial Instrument* and *Person*) and produces two KBLaM units: $(\textit{Financial Instrument}, \textit{regulatory requirement}, V_1)$ and $(\textit{Person}, \textit{regulatory criterion}, V_2)$, where V_1, V_2 are short, evidence-grounded descriptions. Table I illustrates the intermediate artifacts produced by each pipeline stage.

IV. Methodology

This section describes the methodology used to adapt KBLaM to real-world data. The proposed approach comprises dataset selection, preprocessing, model training, and evaluation. The whole pipeline is depicted in 1

A. Preprocessing and Knowledge Extraction Pipeline

To transform data D into format K, a multi-stage preprocessing pipeline was developed. All LLM-based extraction steps use Qwen 3-32B with the recommended generation parameters (temperature 0.6, top-p 0.95, top-k 20, min-p 0). Structured output is enforced via format schemas passed in the system prompt.

1) Running Example of the Transform: To make the pipeline outputs explicit, we provide an end-to-end example showing how a single passage-question pair is converted into KBLaM-compatible knowledge units.

TABLE I. RUNNING EXAMPLE ILLUSTRATING HOW A REGULATORY QUESTION-ANSWER ITEM IS TRANSFORMED INTO KBLAM-COMPATIBLE KNOWLEDGE UNITS. THE EXAMPLE HIGHLIGHTS THE KEY STEPS OF THE PIPELINE: IDENTIFYING THE RELEVANT ENTITIES, DECOMPOSING THE ORIGINAL QUESTION INTO ENTITY-SCOPED SUB-QUESTIONS, AND CONVERTING THE EXTRACTED EVIDENCE INTO STRUCTURED TRIPLES.

Stage	**Output (excerpt)**
Input (Q, P)	**Q:** Under Rules 7.3.2 and 7.3.3, what maturity conditions trigger disclosure? **P:** "...on its maturity will confer (1) an unconditional right to acquire... or (2) the discretion as to his right..."
Answer (optional)	Two conditions: (1) unconditional right to acquire; (2) discretion as to the right to acquire.
Entities	*Financial Instrument*, *Person*
Per-entity decomposition	For *Person*: reformulate to an entity-scoped question ("Under what conditions is a Person considered to hold...?") with the corresponding entity-scoped answer; similarly for *Financial Instrument*.
KBLaM triples K	$(\textit{Financial Instrument}, \textit{regulatory requirement}, V_1)$; $(\textit{Person}, \textit{regulatory criterion}, V_2)$, where V_1, V_2 are evidence-grounded descriptions derived from P.

2) Answer Generation from Context (if needed): For datasets that do not provide gold answers (e.g., ObliQA), an auxiliary LLM (Qwen 3-32B) was used to generate reference answers from the passage-question pairs. These generated answers serve as reference outputs and provide additional textual material for subsequent entity extraction. This step is skipped when gold answers are already available (MultiRC, DROP).

3) Entity Type Discovery: Before extracting concrete entities, the pipeline first identifies the relevant entity types for each question-answer pair. The LLM is prompted to analyze the question and answer and return a list of domain-relevant entity types (e.g., *organization*, *regulation*, *obligation*, *person*) while avoiding generic categories such as "other" or "unknown". Redundant or overlapping types are explicitly suppressed in the prompt (e.g., if both "company" and "organization" are detected, only one is retained). This step adapts the extraction schema to each example rather than relying on a fixed ontology.

4) Entity Extraction and Question Decomposition: Given the discovered entity types, the pipeline extracts concrete entities from the question, answer, and passage. For each identified entity, three attributes are extracted: (1) *entity_name* - the canonical name of the entity; (2) *entity_description_type* - the category of the description (e.g., purpose, regulatory requirement, role); and (3) *entity_description* - a comprehensive description of the entity's attributes and activities derived from the passage.

When a question-answer pair involves multiple entities, the pipeline decomposes it into per-entity sub-pairs: for each entity, a reformulated *question_to_entity* and corresponding *answer_to_entity* are generated, each scoped to a single entity. If only one entity is present, the original question and answer are preserved without reformulation. This decomposition ensures that each resulting knowledge unit captures exactly one entity and its associated information.

5) KBLaM-Compatible Representation: The extracted entity information is mapped to the KBLaM triple format (N_j, R_j, V_j): the entity name becomes N_j, the description type becomes R_j, and the entity description becomes V_j. Each triple thus encodes a single entity and one facet of its role in the source passage. This step produces the final knowledge base K that is embedded into the model's attention mechanism as continuous key-value pairs.

In practice, KBLaM consumes each unit as a *key-value* pair. We construct the key as a short canonical phrase (e.g., "the $\{R_j\}$ of $\{N_j\}$") and use V_j as the value text. This keeps keys lexically compact while preserving the full evidence-grounded content in the value. Multiple triples may be emitted for the same entity when the passage supports multiple facets (e.g., role, obligation, definition), each becoming a separate key–value memory slot.

V. EXPERIMENTAL STUDY

This section presents the results of adapting KBLaM to real-world data using the proposed preprocessing pipeline and evaluates its effectiveness on knowledge-intensive question-answering tasks.

A. Datasets

Three datasets were selected for the experimental study.

ObliQA (Obligation-based Question Answering [11]) contains regulatory and legal texts accompanied by knowledge-intensive questions. This dataset reflects real-world characteristics such as long passages, implicit constraints, and domain-specific terminology, making it suitable for evaluating the proposed approach under realistic conditions.

MultiRC (Multi-Sentence Reading Comprehension [12]) consists of explanatory multi-paragraph passages paired with questions whose correct responses often require aggregating evidence from several sentences rather than extracting a single span. This property makes MultiRC particularly appropriate for assessing the ability of our pipeline to decompose extended textual context into structured knowledge units and to support multi-fact reasoning.

DROP (Discrete Reasoning Over Paragraphs [13]) is designed to evaluate numerical and logical reasoning over natural language text. DROP includes questions that require arithmetic operations, counting, comparison, and other discrete reasoning steps grounded in paragraph-level context. Incorporating DROP enables evaluation of whether externally embedded structured knowledge improves not only semantic retrieval but also compositional and quantitative reasoning.

B. Experimental Setup

The training and evaluation procedure followed the methodology described in the original KBLaM paper, using the Phi-3-mini-4k-instruct model as the base architecture. The preprocessing pipeline was applied to transform the original data into a KBLaM-compatible format by extracting entities, relationships, and structured knowledge representations.

Model performance was evaluated by comparing a baseline (clean) model and a KBLaM-trained model on a subset of 50 test examples using an LLM-based evaluator. To reduce evaluation noise and hallucination effects from the scoring model, each example was evaluated over 50 runs and average scores were reported. The same methodology was applied to assess the quality of the extracted information. To additionally assess entity-type extraction quality, we manually evaluated extracted types on a random sample of 100 examples per dataset and report Precision, Recall, F1-score, and exact-match accuracy of the extracted type set.

C. Evaluation Metrics

In addition to standard precision, recall, and F1-score, we employ three LLM-assessed criteria that capture aspects of answer quality not fully reflected by exact-match metrics:

- **Correctness** $\in [0, 1]$: the degree to which the answer is factually correct (1.0 = fully correct).
- **Coverage** $\in [0, 1]$: the proportion of relevant details from the reference that appear in the answer (1.0 = all details mentioned).
- **Clarity** $\in [0, 1]$: the structural and logical coherence of the answer (1.0 = fully coherent and well-organized).

D. Quantitative Results

On the preprocessed test sets, the KBLaM-trained model achieved the following token-level metrics:

TABLE II. TEST SET METRICS FOR THE KBLAM-TRAINED MODEL

Dataset	Precision	Recall	F1-Score
ObliQA	0.55	0.63	0.58
MultiRC	0.49	0.63	0.55
DROP	0.36	0.50	0.41

These results indicate the model's ability to retrieve and utilize relevant knowledge encoded through the KBLaM mechanism. The comparison between the baseline and trained models according to the LLM-assessed metrics is presented in Table III.

TABLE III. LLM-ASSESSED QUALITY: BASELINE VS. KBLAM-TRAINED MODEL

Model	Correctness	Coverage	Clarity
Baseline (ObliQA)	0.42	0.18	0.65
Trained (ObliQA)	0.62 (+0.20)	0.35 (+0.17)	0.86 (+0.21)
Baseline (MultiRC)	0.15	0.09	0.72
Trained (MultiRC)	0.20 (+0.05)	0.14 (+0.06)	0.93 (+0.21)
Baseline (DROP)	0.15	0.13	0.73
Trained (DROP)	0.19 (+0.04)	0.21 (+0.08)	0.96 (+0.23)

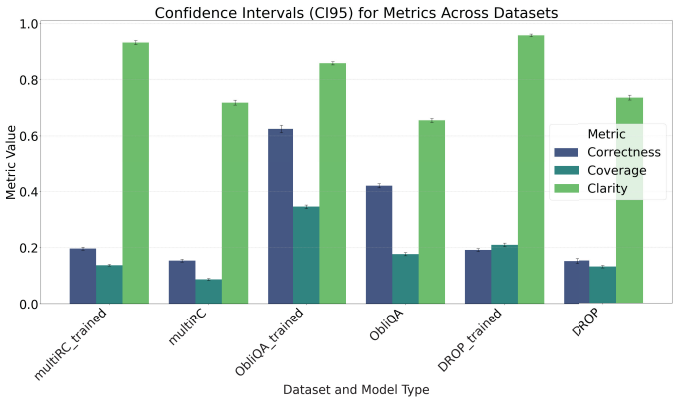


Fig. 2. Comparison of baseline and KBLaM-trained models across LLM-assessed answer-quality metrics. Error bars show 95% confidence intervals over repeated judge runs. The largest gains are observed on ObliQA, while MultiRC and DROP show smaller but consistent improvements, especially in clar-ity.

The trained model demonstrates consistent improvement across all evaluation metrics, indicating more effective utilization of structured knowledge. The largest gains are observed on ObliQA, where the structurally explicit nature of regulatory text aligns well with the pipeline's extraction capabilities.

E. Qualitative Analysis and Extraction Quality

Qualitative examples further illustrate the benefits of KBLaM adaptation. For complex regulatory questions, the baseline model often produced incomplete or truncated answers, failing to capture essential legal conditions or omitting key constraints explicitly stated in the source documents. In contrast, the trained model generated more concise and contextually grounded responses that better reflected the underlying regulatory logic encoded during preprocessing. This suggests that integrating structured knowledge into the attention mechanism improves not only factual retrieval but also answer organization.

However, a limitation becomes apparent when addressing tasks that require not only extraction of relevant knowledge but also compositional reasoning to synthesize a complete response. In such cases, performance gains are less pronounced, indicating that while KBLaM enhances knowledge accessibility, it does not fully resolve challenges associated with higher-order reasoning and inference.

To investigate this observation further, we assessed the quality of the extracted knowledge units themselves using an LLM-based evaluator with the following criteria: **Correctness** (factual accuracy of the extracted unit), **Coverage** (proportion of relevant details captured), and **Importance** (relevance of the entity and its description to the question; 1.0 = critically important). Results are presented in Table IV.

The results indicate that the pipeline performs strongly on structurally explicit regulatory text (ObliQA), achieves moderate effectiveness on multi-sentence reasoning tasks (MultiRC),

TABLE IV. QUALITY OF EXTRACTED KNOWLEDGE UNITS

Dataset	Correctness	Coverage	Importance
ObliQA	0.92	0.93	0.86
MultiRC	0.76	0.64	0.73
DROP	0.38	0.37	0.40

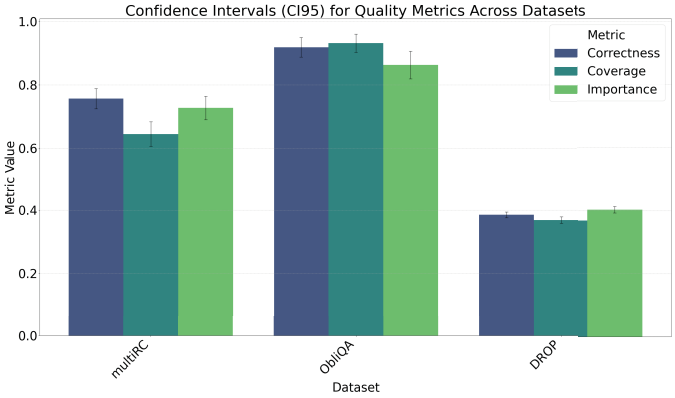


Fig. 3. LLM-based assessment of extracted knowledge-unit quality across datasets, with 95% confidence intervals. ObliQA shows the strongest extraction performance, MultiRC remains moderate, and DROP is substantially more challenging, indi-cating that downstream KBLaM performance is sensitive to the fidelity of the extracted structured knowledge.

and shows limited extraction quality on numerically intensive reasoning datasets (DROP), highlighting its dependence on the structural nature of the source text.

We further analyzed the distribution of extracted entity types in order to better understand why the proposed pipeline performs differently across datasets. Beyond overall extraction quality, the type composition of the resulting knowledge units is important because it determines how naturally the source information can be mapped into KBLaM-style key-value memories. In addition to the distribution plots, we computed the most frequent extracted entity categories in each training set. In ObliQA, the dominant categories were *organization* (6283), *person* (2744), *Regulator* (2030), *Authorised Person* (2029), and *regulatory authority* (1243). In MultiRC, the most frequent categories were *person* (1992), *organization* (602), *location* (380), *concept* (242), and *country* (172). In DROP, the extracted types were dominated by *person* (22264), *number* (6014), *organization* (5003), *percentage* (3825), and *quantity* (3756), highlighting the strong numerical component of this dataset.

As shown in Figure 4a, ObliQA exhibits a more domain-specific and semantically structured distribution, including specialized categories such as *Regulator*, *Authorised Person*, and *regulatory authority*. This is favorable for the proposed pipeline because such categories correspond naturally to stable entity-role descriptions and can therefore be converted into compact KBLaM knowledge units with relatively limited ambiguity.

In contrast, Figure 4b shows that MultiRC has a more heterogeneous distribution dominated by general semantic categories such as *person*, *organization*, and *location*. This reflects the descriptive, narrative nature of the dataset. As a result, the extraction process remains feasible, but the resulting knowledge units are often less domain-constrained and therefore somewhat less precise than in ObliQA.

Finally, Figure 4c indicates that DROP contains a large proportion of numerical and quantitative entity types, including *number*, *percentage*, and *quantity*. This distribution is consistent with the dataset's emphasis on discrete and arithmetic reasoning. Such information is more difficult to encode as stable key–value memories, since successful answering often depends not only on retrieving facts but also on composing or calculating over them, which helps explain the weaker downstream gains observed on DROP.

Overall, these observations suggest that the success of KBLaM adaptation is strongly influenced by the alignment between dataset structure and the assumed knowledge representation format. Datasets with well-defined, domain-specific entity types (such as ObliQA) are more suitable for structured knowledge encoding, while datasets with high heterogeneity or a focus on numerical reasoning present additional challenges.

To further support these observations, we manually evaluated the quality of entity-type extraction. Unlike the previous LLM-based assessment of extracted knowledge units, this analysis directly measures whether the pipeline identifies the correct set of entity types for each example. We report Precision, Recall, F1-score, and exact-match accuracy in Table V.

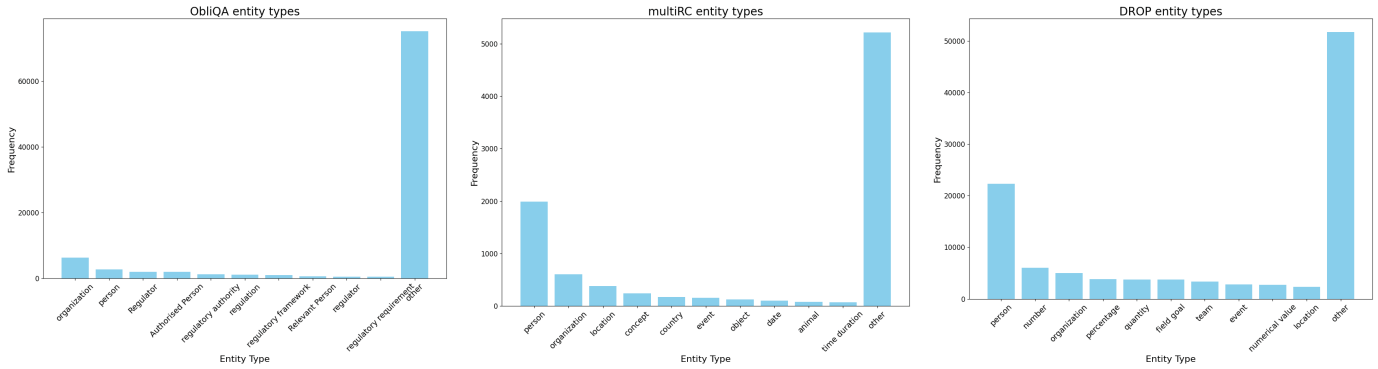
TABLE V. QUALITY OF EXTRACTED ENTITY TYPES BASED ON MANUAL EVALUATION OF 100 EXAMPLES PER DATASET

Dataset	Precision	Recall	F1-score	Accuracy
ObliQA	0.97	0.98	0.98	0.96
MultiRC	0.93	0.91	0.92	0.85
DROP	0.99	0.66	0.79	0.66

For ObliQA, entity type extraction achieves near-perfect performance (F1 = 0.98, Accuracy = 0.96), which is consistent with its well-defined and domain-specific ontology. The high precision and recall indicate that entities are both clearly identifiable and consistently categorized, reinforcing the suitability of this dataset for structured knowledge encoding within the KBLaM framework.

In MultiRC, performance remains high but slightly lower (F1 = 0.92, Accuracy = 0.85), reflecting the more heterogeneous and less formally structured nature of the text.

In contrast, DROP exhibits a distinct pattern: extremely high precision (0.99) but substantially lower recall (0.66), resulting in a moderate F1-score (0.79) and lower overall accuracy (0.66). This indicates that while the extracted entity types are highly accurate when detected, a significant portion of relevant entities is not captured. This behavior aligns with the dataset's emphasis on numerical and implicit quantities, which are often harder to detect and classify using standard entity extraction approaches.



(a) ObliQA (b) MultiRC (c) DROP

Fig. 4. Distribution of extracted entity types in the three training sets. ObliQA exhibits a domain-specific regulatory schema, MultiRC shows a broader narrative-oriented distribution, and DROP is dominated by numerical and quantitative categories, illustrating important structural differences that affect the suitability of each dataset for KBLaM-style structured knowledge encoding.

Overall, these results highlight that entity extraction quality is strongly influenced by the structural properties of the dataset. High recall and balanced performance, as seen in ObliQA, are critical for effective knowledge construction, whereas recall limitations, as in DROP, can lead to incomplete knowledge representations and reduced downstream performance.

F. Summary of Results

Overall, the experimental results demonstrate that the proposed preprocessing pipeline enables effective adaptation of KBLaM to real-world data. The trained model shows improved correctness, broader coverage of key details, and higher clarity compared to the baseline, validating the feasibility of applying KBLaM to complex, real-world knowledge bases. At the same time, the extraction quality analysis reveals that downstream QA performance is closely tied to the fidelity of the knowledge extraction step, with the largest gains observed where the pipeline can produce high-quality structured units.

VI. Ablation Study

To evaluate the contribution of structured knowledge descriptions within the KBLaM-based pipeline, we performed an ablation study on the MultiRC dataset with two controlled modifications: (1) *shuffled descriptions*, where descriptions were randomly reassigned across entities, breaking semantic alignment while preserving lexical content; and (2) *empty descriptions*, where description fields were removed entirely, effectively disabling knowledge injection.

The quantitative results are presented in Table VI.

TABLE VI. Ablation study results on MultiRC

Configuration	Precision	Recall	F1-Score
Baseline	0.49	0.63	0.55
Shuffled	0.48	0.58	0.52
Empty	0.46	0.57	0.51

Performance progressively decreases from the baseline to the shuffled and empty-description settings. The shuffled configuration results in a noticeable drop across all metrics, indicating that semantic misalignment between keys and descriptions weakens retrieval and reasoning. The empty-description condition leads to the largest degradation, confirming that the presence of structured knowledge representations significantly contributes to overall performance.

These findings demonstrate that the improvements achieved by the proposed approach stem from meaningful and correctly aligned knowledge encoding rather than from incidental increases in input length or architectural complexity.

Beyond the internal ablations above, we also report a lightweight retrieval-augmented generation (RAG) reference as an external point of comparison. While the shuffled/empty settings isolate the role of *semantic alignment* in KBLaM's injected knowledge, this RAG reference estimates the quality achievable by standard test-time retrieval without any KB-side memory integration. Specifically, on the same 50-question subset, we embed passages with `all-MiniLM-L6-v2`, retrieve the top-20 passages per question, and generate answers using `Qwen3` served via a vLLM OpenAI-compatible API. Using the same LLM-as-judge rubric with a single evaluation run per example, the RAG setup achieves Correctness 0.73, Coverage 0.66, and Clarity 0.87. We use this number set only to contextualize retrieval-only behavior under our evaluation protocol.

Taken together, the high *Clarity* but only partial *Coverage* suggests that retrieval-time access to evidence produces well-structured answers, yet still leaves a substantial fraction of gold key points uncovered (0.66). We therefore treat this RAG score as a practical reference for the remaining headroom in information preservation and aggregation when converting documents into KBLaM knowledge units.

VII. Discussion

The experimental results reveal a consistent pattern: downstream QA performance is strongly correlated with the quality of the upstream knowledge extraction step. On ObliQA, where the pipeline achieves high extraction correctness (0.92) and coverage (0.93), the KBLaM-trained model yields the largest gains in answer correctness (+0.20) and coverage (+0.17). On MultiRC, moderate extraction quality (correctness 0.76, coverage 0.64) translates into more modest improvements. On DROP, where extraction scores fall below 0.40 across all criteria, QA gains are marginal. This gradient suggests that the primary bottleneck for real-world KBLaM deployment is not the model's attention-based integration mechanism itself, but rather the fidelity of the structured knowledge supplied to it.

A notable exception to this pattern is the clarity metric, which improves substantially across all three datasets (+0.21 to +0.23) regardless of extraction quality. One possible explanation is that the KBLaM attention mechanism imposes an implicit structural prior on generation: even when the injected knowledge is noisy or incomplete, the model learns to organize its output around the provided key-value structure, producing more coherent answers. This finding suggests that structured knowledge injection may confer organizational benefits beyond pure factual recall.

The ablation study on MultiRC provides further evidence that the observed gains are not artifacts of increased input length. The shuffled-description condition preserves the same token count as the baseline yet degrades performance, confirming that semantic alignment between keys and values is essential. The empty-description condition removes knowledge content entirely and produces the largest drop, ruling out the possibility that the key embeddings alone are sufficient.

Several limitations of the current study should be acknowledged. First, the preprocessing pipeline relies on an auxiliary LLM (Qwen 3.0) for answer generation and entity extraction, which introduces a dependency on the quality and availability of a capable language model at data preparation time. The extraction quality results on DROP indicate that this LLM-based approach struggles with passages that require numerical reasoning or contain densely packed quantitative information, suggesting that specialized extraction strategies may be needed for such domains. Second, the evaluation was conducted on subsets of 50 examples per dataset, which, while mitigated by averaging over 50 evaluation runs per example, limits the statistical power of the reported comparisons. Third, the use of an LLM-based evaluator for correctness, coverage, and clarity introduces its own potential biases; although we reduce variance through repeated scoring, systematic biases of the scoring model cannot be fully excluded.

From a broader perspective, these results highlight the importance of data-centric approaches in knowledge-augmented language modeling. While much of the existing literature focuses on architectural innovations for knowledge integration, our findings indicate that the quality of the input knowledge representation may have a greater impact on end-task performance than the integration mechanism itself. This observation aligns with recent trends in the machine learning community emphasizing data quality over model complexity and suggests that future work on KBLaM and similar frameworks should invest equally in robust knowledge extraction pipelines.

VIII. Conclusion

In this work, we investigated the problem of adapting KBLaM to real-world data, which is characterized by noise, complexity, and heterogeneous structure.

A dedicated preprocessing pipeline was designed to transform complex, unstructured texts into KBLaM-compatible knowledge representations. Using this pipeline, a KBLaM-augmented model was trained following the methodology of the original KBLaM framework. Experimental evaluation on three benchmarks demonstrated that the trained model consistently outperforms the baseline in correctness, coverage of key details, and clarity of generated answers.

Qualitative analysis showed that the adapted model can effectively leverage embedded structured knowledge when answering complex questions, producing more complete and relevant responses than the baseline. However, extraction quality analysis also revealed that the pipeline's effectiveness depends on the structural nature of the source text: gains are most pronounced for structurally explicit documents and diminish for tasks requiring numerical or discrete reasoning, where the extraction step itself becomes the bottleneck.

An ablation study further validated the approach. By comparing the baseline configuration with variants using shuffled and empty knowledge descriptions, we observed consistent degradation in precision, recall, and F1-score when semantic alignment between keys and descriptions was disrupted or removed. These results confirm that the observed improvements are directly attributable to structured and semantically coherent knowledge encoding rather than to incidental architectural modifications or increased input length.

Future work will focus on scaling the approach to larger and more diverse knowledge bases, improving robustness to noisy entity extraction, and exploring dynamic knowledge updates without retraining, further advancing the applicability of KBLaM in real-world systems.

Acknowledgment

This work was supported by the Russian Science Foundation, agreement no. 24-71-00115, https://rscf.ru/en/project/24-71-00115/.

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[2] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense passage retrieval for open-domain question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 6769–6781. [Online]. Available: https://aclanthology.org/2020.emnlp-main.550/

[3] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "Retrieval augmented language model pre-training," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 3929–3938. [Online]. Available: https://proceedings.mlr.press/v119/guu20a.html

[4] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. B. Van Den Driessche, J.-B. Lespiau, B. Damoc, A. Clark, D. De Las Casas, A. Guy, J. Menick, R. Ring, T. Hennigan, S. Huang, L. Maggiore, C. Jones, A. Cassirer, A. Brock, M. Paganini, G. Irving, O. Vinyals, S. Osindero, K. Simonyan, J. Rae, E. Elsen, and L. Sifre, "Improving language models by retrieving from trillions of tokens," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 17–23 Jul 2022, pp. 2206–2240. [Online]. Available: https://proceedings.mlr.press/v162/borgeaud22a.html

[5] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021. [Online]. Available: https://arxiv.org/abs/2106.09685

[6] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: efficient finetuning of quantized llms," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2023.

[7] X. Wang, L. Mikaelyan, T. Isazawa, and J. Hensman, "Kblam: Knowledge base augmented language model," in *International Conference on Learning Representations (ICLR)*, 2025. [Online]. Available: https://openreview.net/forum?id=aLsMzkTej9

[8] H. Huang, H. T. Tsang, J. Bai, X. Peng, G. Zhang, and Y. Song, "Atlaskv: Augmenting llms with billion-scale knowledge graphs in 20gb vram," in *International Conference on Learning Representations (ICLR)*, 2026. [Online]. Available: https://openreview.net/forum?id=6i1jVAYbHs

[9] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, S. Kirrane, J. E. Labra Gayo, R. Navigli, S. Neumaier, A. Ngonga Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, and A. Zimmermann, "Knowledge graphs," *ACM Computing Surveys*, vol. 54, no. 4, 2021. [Online]. Available: https://dl.acm.org/doi/10.1145/3447772

[10] L. Zhong, J. Wu, Q. Li, H. Peng, and X. Wu, "A comprehensive survey on automatic knowledge graph construction," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–62, 2023. [Online]. Available: https://dl.acm.org/doi/10.1145/3618295

[11] T. Gokhan, K. Wang, I. Gurevych, and T. Briscoe, "Rirag: Regulatory information retrieval and answer generation," 2024. [Online]. Available: https://arxiv.org/abs/2409.05677

[12] D. Khashabi, S. Chaturvedi, M. Roth, S. Upadhyay, and D. Roth, "Looking beyond the surface:a challenge set for reading comprehension over multiple sentences," in *Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018.

[13] D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, and M. Gardner, "DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs," in *Proc. of NAACL*, 2019.