ICLR

AtlasKV

Augmenting LLMs with Billion-Scale Knowledge Graphs in 20GB VRAM

Haoyu Huang[1], Hong Ting Tsang[1], Jiaxin Bai[1], Xi Peng[2], Gong Zhang[2], Yangqiu Song[1]

[1] CSE,HKUST [2] Theory Lab,Huawei

Rio de Janeiro, Brazil

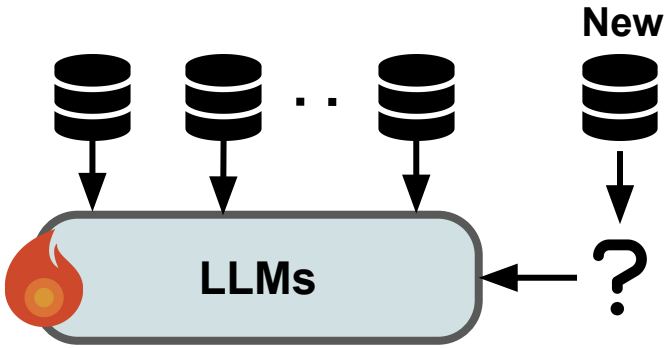
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

THE DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING 計算機科學及工程學系

KnowComp Group Understanding the World by Computational Knowledge

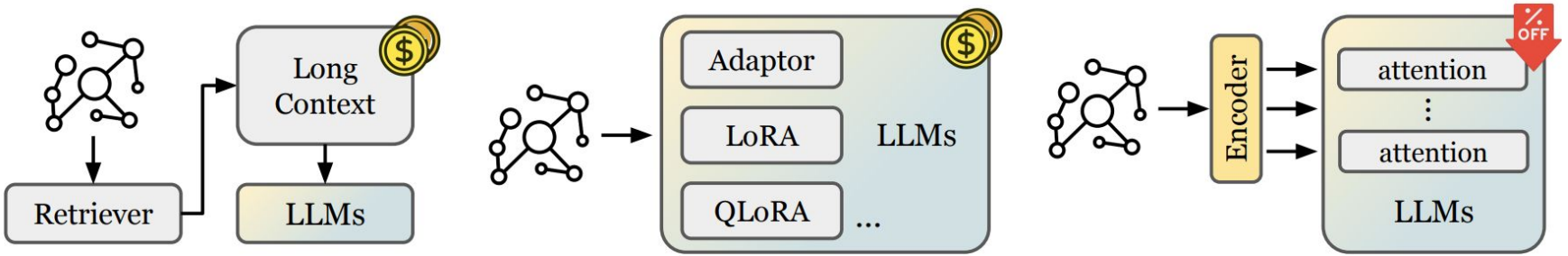
HUAWEI

Now What's Wrong with LLMs?



- ***Expensive*** to train with full parameters.
- Hard to quickly adapt to ***new knowledge***.

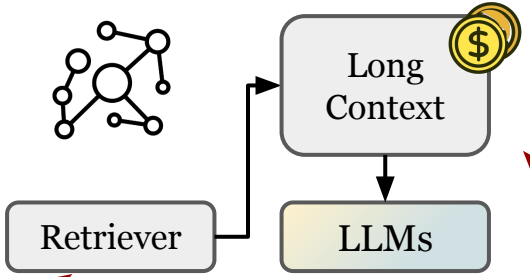
Knowledge Augmentation Paradigms for LLMs



(a) Non-Parametric Methods (b) Traditional Parametric Methods (c) AtlasKV

Knowledge Augmentation Paradigms for LLMs

Without parameter modifications in LLMs.



(a) Non-Parametric Methods

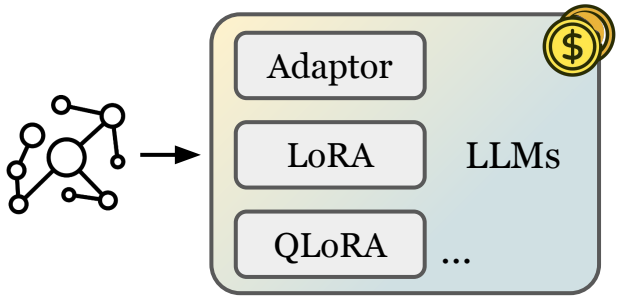
But

(1) Expensive searches (e.g. nearest-neighbor searches).
(2) Performance limited by the retriever.

(1) Long context prior introduces substantial inference latency, especially when augment with very large scale knowledge.
(2) Too much token cost.

Examples: RAG, Graph-based RAG, Agentic RAG, etc.

Knowledge Augmentation Paradigms for LLMs



(b) Traditional Parametric Methods

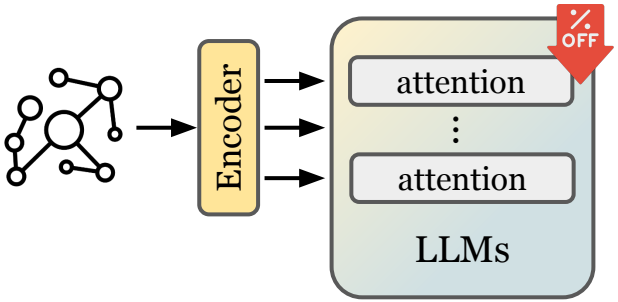
Do not need any external retriever or long context prior.

But

Require re-training the model when adapting to new knowledge.
Also expensive.

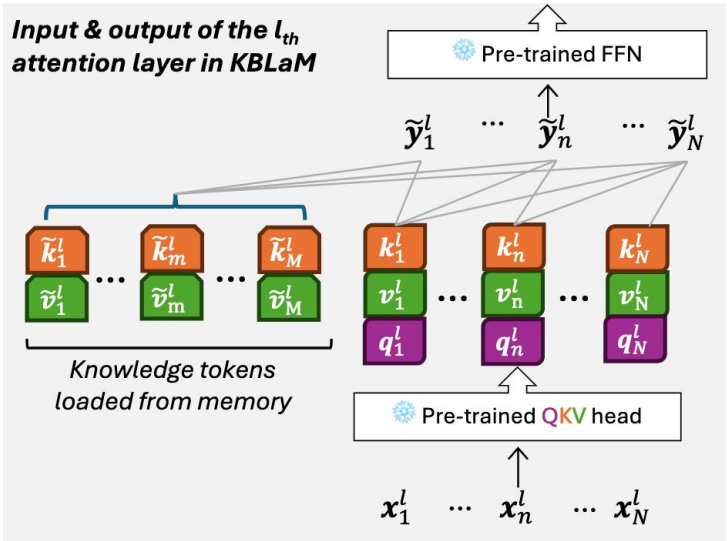
Examples: Adaptor, LoRA, QLoRA, Prefix-Tuning, Prompt-Tuning, etc.

Knowledge Augmentation Paradigms for LLMs



(c) AtlasKV

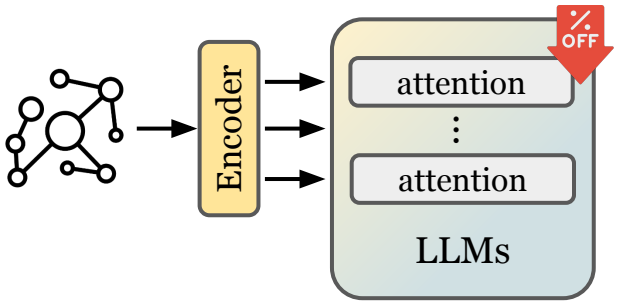
Examples: KBLaM.



(1) Do not need any external retriever or long context prior.
(2) Could adapt to new knowledge without training.

$$\tilde{\mathbf{y}}_n^{(l)} = \frac{\sum_{m=1}^{M} \exp(\langle \tilde{\mathbf{q}}_n^{(l)}, \tilde{\mathbf{k}}^{(l)m} \rangle / \sqrt{D}) \tilde{\mathbf{v}}^{(l)m}}{\sum_{m=1}^{M} \exp(\langle \tilde{\mathbf{q}}_n^{(l)}, \tilde{\mathbf{k}}^{(l)m} \rangle / \sqrt{D})} + \frac{\sum_{i=1}^{N} \exp(\langle \mathbf{q}_n^{(l)}, \mathbf{k}^{(l)i} \rangle / \sqrt{D}) \mathbf{v}^{(l)i}}{\sum_{i=1}^{N} \exp(\langle \mathbf{q}_n^{(l)}, \mathbf{k}^{(l)i} \rangle / \sqrt{D})},$$

What Limits the Paradigm of KBLaM?



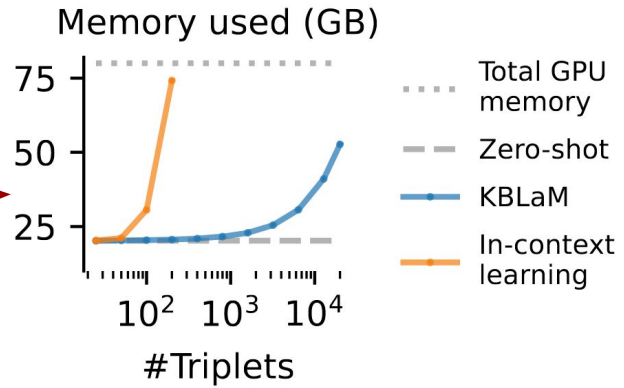
Limited Enquiry Attributes

Tell me the *purpose* of ...
Describe the *objectives* of ...
What's the *description* of ...

Diverse Enquiry Attributes

What's the *record-breaker* of ...
Tell me the *co-author* of ...
What's the *rank* of ...
Describe the *cause* of ...
...

(1) Lack of high quality Q-K-V training data.
(2) Poor scalability. Too much cost when scale external knowledge to very large.



Solution: AtlasKV

A **scalable**, **effective**, and **general** way to augment LLMs with **billion-scale** knowledge graphs (KGs) (e.g. 1B triples) using **very little GPU memory cost** (e.g. less than 20GB VRAM).

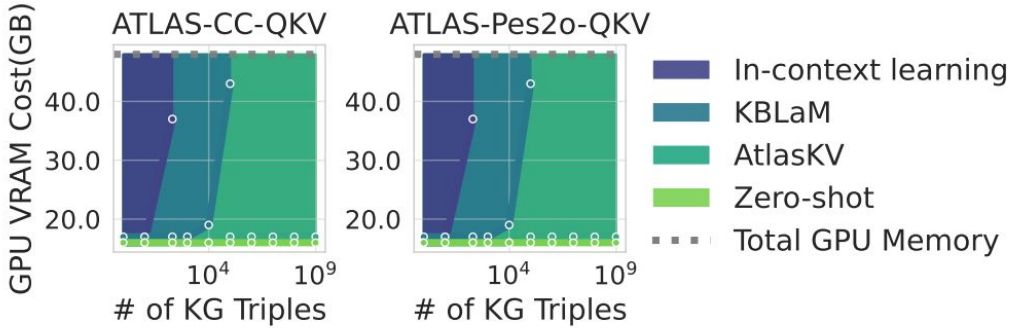
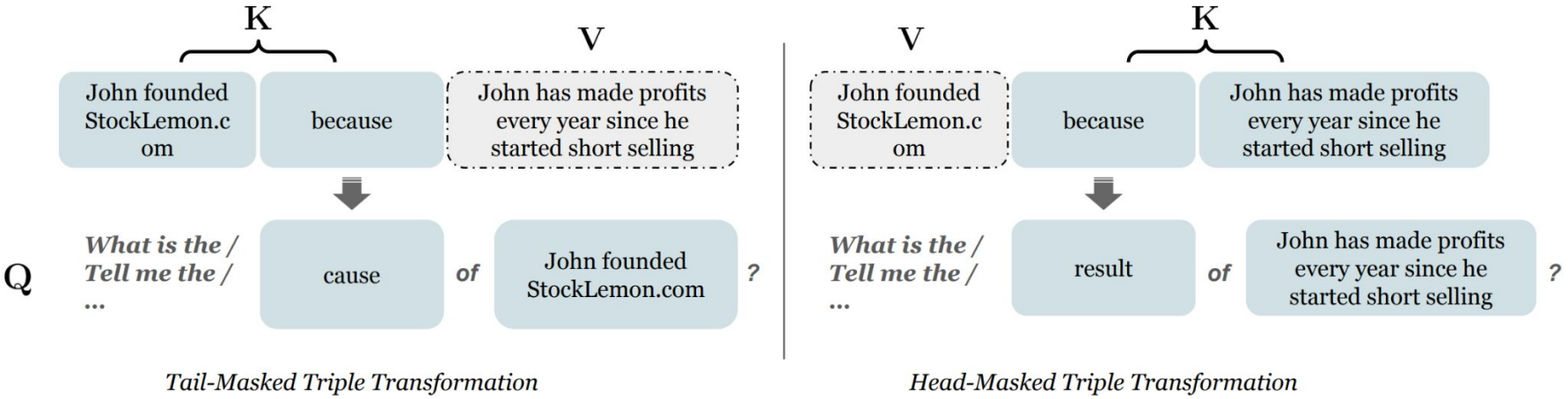


Figure 4: GPU memory usage comparison of AtlasKV and other methods across various KG sizes from 1 to 1B triples.

AtlasKV: Why Effective and General?

KG2KV, a new KG integration paradigm that naturally converts KG triples into Q-K-V data, enabling LLMs to achieve both enhanced generalization performance and effective knowledge integration.



AtlasKV: Why Effective and General?

Compared with fully synthetic method in KBLaM, **more diverse training data** and **lower token cost**.

Table 1: Comparison of the data diversity ratio and average token cost between KG2KV and synthetic method.

Method	Diversity Ratio ↑	Avg. Token Cost ↓
Synthetic	0.003%	349.9
KG2KV	**7.864%**	**165.7**

Prompt Template for KG2KV.

System Message:

Task: Convert relation phrase to natural noun based on missing entity position.
Rules:
- **Missing head**: Passive relations → agent nouns ("govern" → "governor", "is participated by" → "participation")
- **Missing tail**: Active relations → object nouns ("produces" → "product", "achieves" → "achievement")

Output: Natural noun only.

Examples:
- ("is participated by", "head") → "participation"
- ("is participated by", "tail") → "participant"
- ("produces", "head") → "producer"
- ("produces", "tail") → "product"

User Message:

relation: {relation}, missing: {missing}

AtlasKV: Why Effective and General?

Some cases.

Q	*K*	*V*
What is the *description* of Elara Moonshadow?	the *description* of Elara Moonshadow	The *description* of Elara Moonshadow is a skilled botanist with a passion for rare plants.
Describe the *description* of Thorne Blackwood?	the *description* of Thorne Blackwood	The *description* of Thorne Blackwood is a renowned chef known for his innovative culinary techniques.
Provide details on the *objectives* of Zara Nightingale?	the *objectives* of Zara Nightingale	The *objectives* of Zara Nightingale is to perform in prestigious concert halls worldwide.
Can you let me know the *purpose* of Lyra Starfire?	the *purpose* of Lyra Starfire	The *purpose* of Lyra Starfire is to preserve marine biodiversity.
Can you explain the *description* of Jaxon Wildheart?	the *description* of Jaxon Wildheart	The *description* of Jaxon Wildheart is a tech entrepreneur with a knack for innovative solutions.
What insights can you provide about the *objectives* of Kaelith Silverwind?	the *objectives* of Kaelith Silverwind	The *objectives* of Kaelith Silverwind is to document endangered animals.

Table 6: Samples from Synthetic dataset. The enquiry attributes have been marked in *italics*.

Q	*K*	*V*
What is the *explanation* of Postsocialist scholars?	the *explanation* of Postsocialist scholars	The *explanation* of Postsocialist scholars is the developments as a backlash against the 'feminizing' nature of the socialist state.
Can you explain the *cause* of World records?	the *cause* of World records	The *cause* of World records is World records in Paralympic powerlifting are ratified by the International Paralympic Committee.
Can you elaborate on the *rank* of Ramble On?	the *rank* of Ramble On	The *rank* of Ramble On is number 5 on the list of the 40 greatest Led Zeppelin songs.
How would you describe the *favorite* of Dick the Mockingbird?	the *favorite* of Dick the Mockingbird	The *favorite* of Dick the Mockingbird is among at least four mockingbirds the president had while in office.
Can you inform me about the *publication* of Hensley?	the *publication* of Hensley	The *publication* of Hensley is Fifty Miles from Tomorrow , a memoir of Alaska and the real people.
Tell me about the *threat* of African coral reefs?	the *threat* of African coral reefs	The *threat* of African coral reefs is industrial run-offs and pollutants, untreated sewage and the increasing sediment flows in rivers.

Table 7: Samples from ATLAS-Wiki-QKV dataset. The enquiry attributes have been marked in *italics*.

AtlasKV: Why Effective and General?

Empirical evidences from training dynamics of AtlasKV.

From a specific training step, the model regularly **start learning to retrieve relevant knowledge from the external KG triples, instead of brute force over-fitting**.

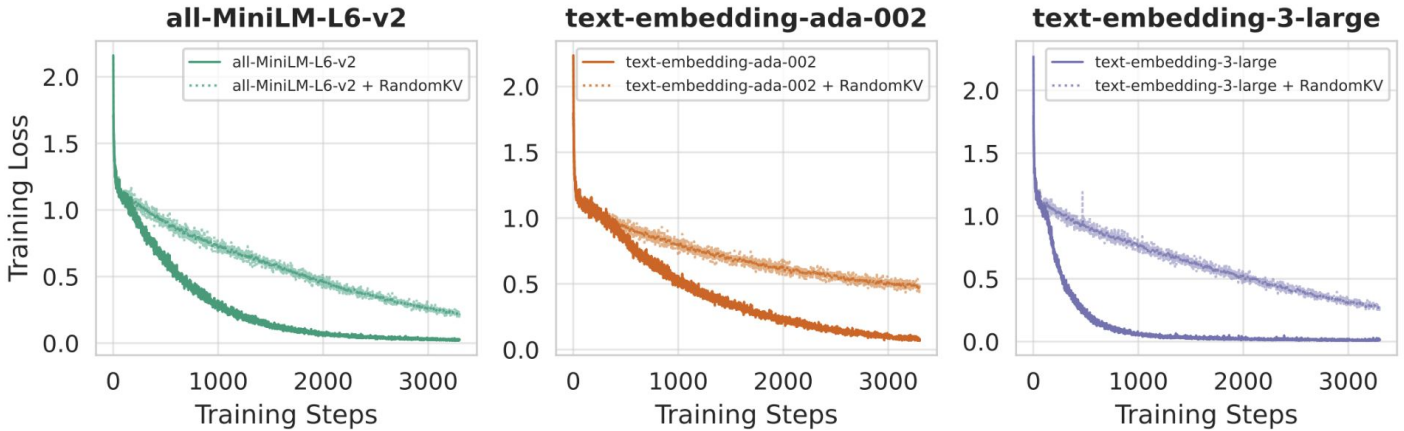
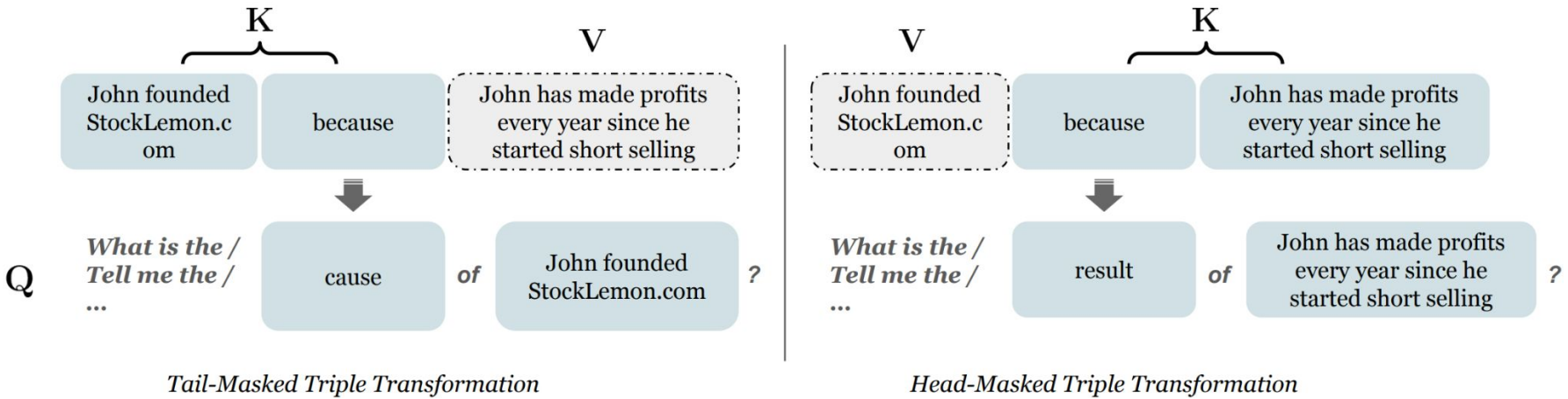


Figure 7: The training loss curves of AtlasKV with correct and random paired key-value embeddings (KGKVs) across three different sentence encoders.

AtlasKV: Why Effective and General?

Note that for the training data, we usually select **named entities as the key to mask**, and select **event entities and relations as the value**.



AtlasKV: Why Scalable?

HiKVP, a hierarchical key-value pruning algorithm that dramatically reduces computational and memory overhead while maintaining high knowledge grounding accuracy during inference time.

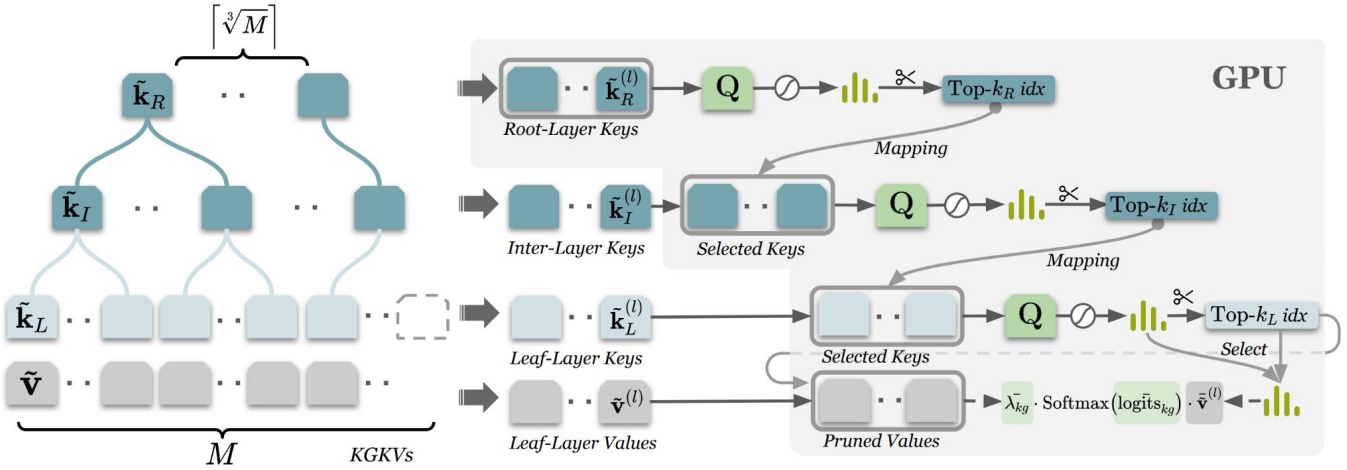


Figure 3: An overview of hierarchical key-value pruning (HiKVP) with three layers of knowledge keys at the l-th attention layer. The gray background indicate that the part is stored and computed in the GPU memory.

AtlasKV: Why Scalable?

HiKVP, a hierarchical key-value pruning algorithm that dramatically reduces computational and memory overhead while maintaining high knowledge grounding accuracy during inference time.

KBLaM:

$$\tilde{\mathbf{y}}_n^{(l)} = \frac{\sum_{m=1}^{M} \exp(\langle \tilde{\mathbf{q}}_n^{(l)}, \tilde{\mathbf{k}}^{(l)m} \rangle / \sqrt{D}) \tilde{\mathbf{v}}^{(l)m} + \sum_{i=1}^{n} \exp(\langle \mathbf{q}_n^{(l)}, \mathbf{k}^{(l)i} \rangle / \sqrt{D}) \mathbf{v}^{(l)i}}{\sum_{m=1}^{M} \exp(\langle \tilde{\mathbf{q}}_n^{(l)}, \tilde{\mathbf{k}}^{(l)m} \rangle / \sqrt{D}) + \sum_{i=1}^{n} \exp(\langle \mathbf{q}_n^{(l)}, \mathbf{k}^{(l)i} \rangle / \sqrt{D})}$$

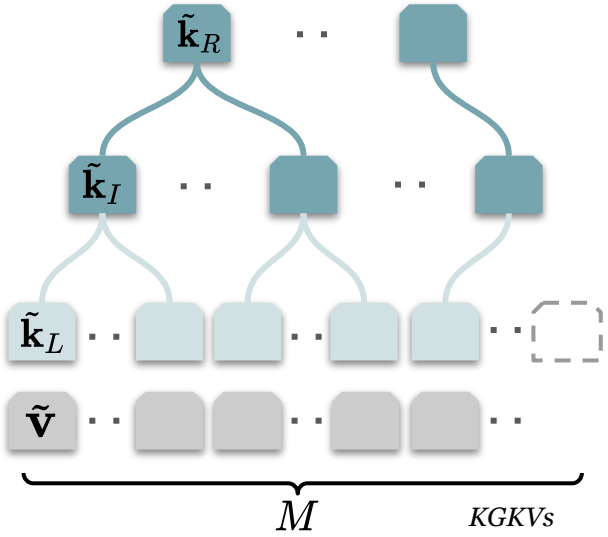
AtlasKV:

$$\tilde{\mathbf{y}}_n^{(l)} = \lambda_{kg} \cdot \text{Softmax}\left(\text{logits}_{kg_L}\right) \cdot \tilde{\mathbf{v}}^{(l)} + \lambda_{seq} \cdot \text{Softmax}\left(\text{logits}_{seq}\right) \cdot \mathbf{v}^{(l)} \quad \text{(train)}$$

$$\tilde{\mathbf{y}}_n^{(l)} = \bar{\lambda_{kg}} \cdot \text{Softmax}\left(\bar{\text{logits}}_{kg}\right) \cdot \bar{\tilde{\mathbf{v}}}^{(l)} + \lambda_{seq} \cdot \text{Softmax}\left(\text{logits}_{seq}\right) \cdot \mathbf{v}^{(l)} \quad \text{(inference)}$$

AtlasKV: Why Scalable?

To share the computational and memory burden equally, we set the size of clusters in each layer to be the same, which is $\left\lceil \sqrt[3]{M} \right\rceil$.



# of Keys	Time Complexity	Memory Complexity
$\left\lceil \left\lceil M^{\frac{2}{3}} \right\rceil M^{-\frac{1}{3}} \right\rceil$	$\mathcal{O}\left(\sqrt[3]{M}D + \sqrt[3]{M}\log(k_R)\right)$	$\mathcal{O}\left(\sqrt[3]{M}(N+D)\right)$
$\left\lceil M^{\frac{2}{3}} \right\rceil$	$\mathcal{O}\left(k_R\sqrt[3]{M}D + k_R\sqrt[3]{M}\log(k_I)\right)$	$\mathcal{O}\left(k_R\sqrt[3]{M}(N+D)\right)$
M	$\mathcal{O}\left(k_I\sqrt[3]{M}D + k_I\sqrt[3]{M}\log(k_L)\right)$	$\mathcal{O}\left(k_I\sqrt[3]{M}(N+D)\right)$
M		

Attention Top-K

AtlasKV: Why Scalable?

Compared with ICL, RAG, KBLaM, AtlasKV is more scalable.

Method	Time Complexity	Memory Complexity
ICL	$\mathcal{O}\left((MT+N)^2 \cdot D\right)$	$\mathcal{O}\left((MT+N)\cdot(MT+N+D)\right)$
RAG	$\mathcal{O}\left(M+RT+(RT+N)^2\right)\cdot D$	$\mathcal{O}\left((RT+N)\cdot(RT+N+D)\right)$
KBLaM	$\mathcal{O}\left((M+N)\cdot N\cdot D\right)$	$\mathcal{O}\left((M+N)\cdot(N+D)\right)$
AtlasKV	$\mathcal{O}\left((C_t\sqrt[3]{M}+N)\cdot N\cdot D\right)$	$\mathcal{O}\left((C_m\sqrt[3]{M}+N)\cdot(N+D)\right)$

Table 2: Comparison of the time and memory complexity of AtlasKV, KBLaM, RAG, and ICL methods, where the parts marked in teal color represent they could be very large.

$$C_t = 1 + k_R + k_I$$

$$C_m = \max\left(1, k_R, k_I\right)$$

AtlasKV: How Effective and General?(Experiments)

AtlasKV is more accurate and generalizable with KGKVs.
Knowledge grounding accuracy:

Method	Steps	10^1 Triples ACC@1	10^1 Triples ACC@5	10^2 Triples ACC@1	10^2 Triples ACC@5	10^3 Triples ACC@1	10^3 Triples ACC@5	10^4 Triples ACC@1	10^4 Triples ACC@5
Eval on Enron									
KBLaM	3e3	**100.0**	**100.0**	50.9	76.4	29.1	56.4	9.1	20.0
	2e4	90.0	**100.0**	50.9	83.6	25.5	47.3	7.3	23.6
AtlasKV (128-64-16)	3e3	**100.0** (+0.0)	**100.0** (+0.0)	67.3 (+16.4)	90.9 (+7.3)	41.8 (+12.7)	50.9	21.8 (+12.7)	32.7 (+12.7)
AtlasKV w/o HiKVP	3e3	**100.0** (+0.0)	**100.0** (+0.0)	**76.4** (+25.5)	**92.7** (+9.1)	**56.4** (+27.3)	**80.0** (+23.6)	**27.3** (+18.2)	**47.3** (+27.3)
Eval on ATLAS-Pes2o-QA									
KBLaM	3e3	40.0	80.0	16.4	45.5	5.5	14.5	0.0	3.6
	2e4	50.0	80.0	25.5	52.7	3.6	14.5	0.0	5.5
AtlasKV (128-64-16)	3e3	90.0 (+40.0)	100.0 (+20.0)	87.3 (+61.8)	92.7 (+40.0)	52.7 (+47.2)	70.9 (+56.4)	16.4 (+16.4)	49.0 (+43.5)
AtlasKV w/o HiKVP	3e3	**100.0** (+50.0)	**100.0** (+20.0)	**92.7** (+67.2)	**100.0** (+47.3)	**72.7** (+67.2)	**90.9** (+76.4)	**47.3** (+47.3)	**67.2** (+61.7)
Eval on ATLAS-CC-QA									
KBLaM	3e3	60.0	90.0	21.8	38.2	12.7	23.6	3.6	10.9
	2e4	50.0	**100.0**	23.6	56.4	10.9	21.8	3.6	10.9
AtlasKV (128-64-16)	3e3	**100.0** (+0.0)	**100.0** (+0.0)	89.1 (+65.5)	90.9 (+34.5)	61.8 (+49.1)	74.5 (+50.9)	40.0 (+36.4)	54.5 (+43.6)
AtlasKV w/o HiKVP	3e3	**100.0** (+0.0)	**100.0** (+0.0)	**96.4** (+72.8)	**100.0** (+43.6)	**83.6** (+70.9)	**96.4** (+72.8)	**61.8** (+58.2)	**81.8** (+70.9)

Table 3: The knowledge grounding performance of AtlasKV against KBLaM with all-MiniLM-L6-v2 as the sentence encoder on three OOD evaluation datasets across various tuning steps and KG sizes. We defaultly set the top-k in HiKVP to 128, 64, and 16 for the k_R, k_I, k_L respectively.

AtlasKV: How Effective and General?

Answer relevance:

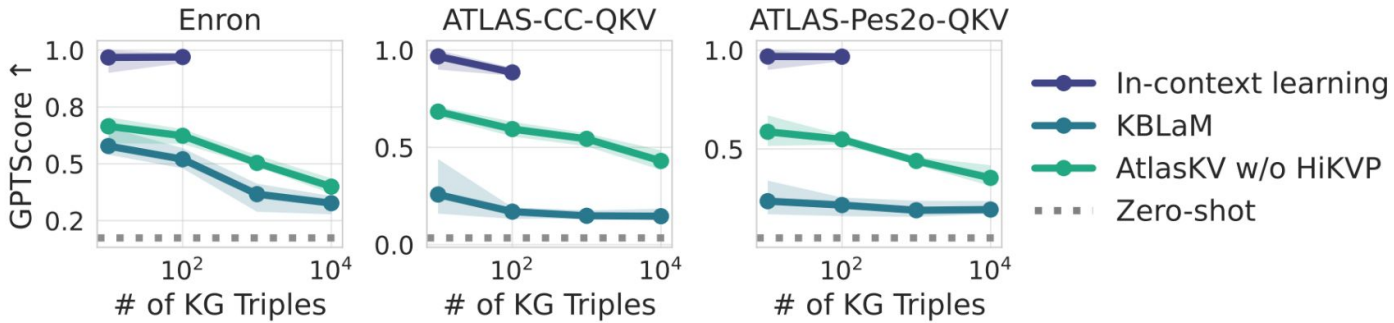


Figure 5: Scored by GPT-4o between 0 and 1, the shaded area exhibits the standard error over 5 random seeds. The score of each random seed is also the average of 5 generation results.

Remarkably, despite having only limited training samples with enquiry attributes similar to Enron in ATLAS-Wiki-QKV, **AtlasKV still outperforms KBLaM** on both knowledge grounding accuracy and answer relevance metrics, **even though KBLaM's training data contains exactly the same enquiry attributes as Enron.**

AtlasKV: How Effective and General?

Ablation Study:

Method	Steps	10^1 Triples ACC@1	10^1 Triples ACC@5	10^2 Triples ACC@1	10^2 Triples ACC@5	10^3 Triples ACC@1	10^3 Triples ACC@5	10^4 Triples ACC@1	10^4 Triples ACC@5
Eval on ATLAS-Pes2o-QA									
AtlasKV w/o HiKVP	3e3	**100.0**	**100.0**	**92.7**	**100.0**	**72.7**	**90.9**	**47.3**	**67.2**
AtlasKV w/o HiKVP & Event	3e3	90.0	**100.0**	80.0	89.1	34.5	63.6	9.1	36.4
AtlasKV w/o HiKVP & Entity	3e3	**100.0**	**100.0**	49.0	67.3	20.0	30.9	3.6	5.5
Eval on Enron									
AtlasKV w/o HiKVP	3e3	**100.0**	**100.0**	**76.4**	**92.7**	**56.4**	**80.0**	**27.3**	**47.3**
AtlasKV w/o HiKVP & Event	3e3	80.0	**100.0**	73.6	84.5	48.0	66.0	10.9	38.2
AtlasKV w/o HiKVP & Entity	3e3	40.0	**100.0**	40.0	74.5	16.4	27.3	1.8	9.1

Table 4: The knowledge grounding performance of different variants of AtlasKV with all-MiniLM-L6-v2 as the sentence encoder on three OOD evaluation datasets across various tuning steps and KG sizes.

Cooperating named and event entities together in KG2KV process helps with the model's learning.

AtlasKV: How Effective and General?

Influence of different Top-K in HiKVP (1e5 triples):

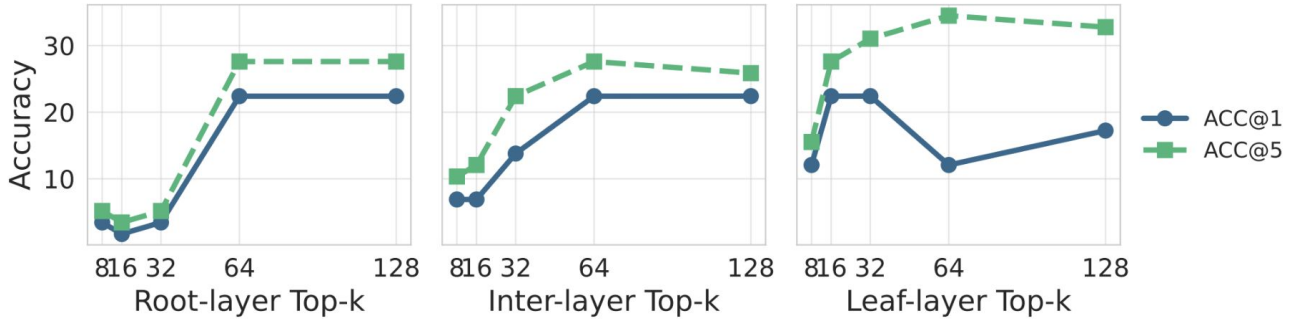


Figure 6: The knowledge grounding accuracy of AtlasKV on ATLAS-CC-QKV with different top-k settings at each layer.

The accurate retrieval ability of AtlasKV is stronger than the fuzzy retrieval ability of it. And the reason why too large k_I or k_L will hurt the performance might be that **the noise candidate keys selected in early attention layers would influence the retrieval accuracy of the later attention layers.**

AtlasKV: How Scalable?

With the increasing of KG scale, the VRAM usage of AtlasKV **even just a little bit higher than the zero-shot generation**.

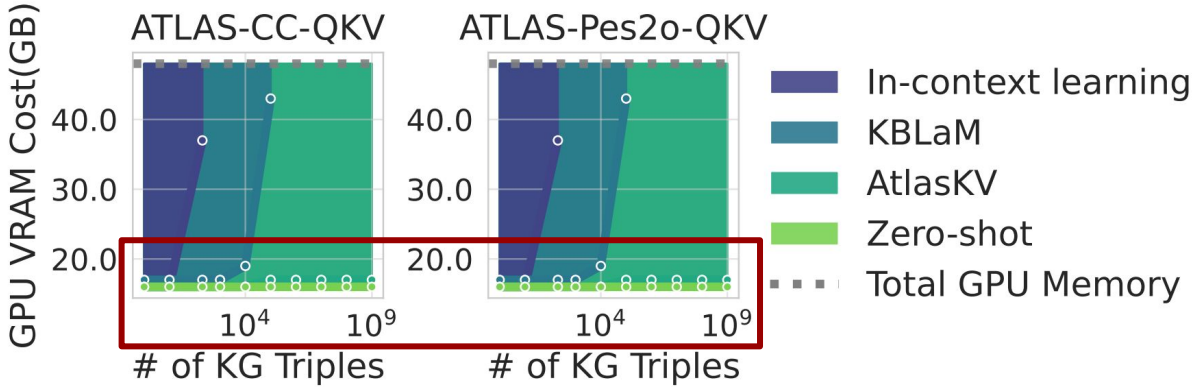


Figure 4: GPU memory usage comparison of AtlasKV and other methods across various KG sizes from 1 to 1B triples.

AtlasKV: Case Study

Sample Outputs.

Relevant Triple: (*MOROCCO*; *consider*; *synthetic biology should be considered as a new and emerging issue*)
Q: *Can you elaborate on the opinion of MOROCCO?*
K: *the opinion of MOROCCO*
V: *The opinion of MOROCCO is synthetic biology should be considered as a new and emerging issue.*

AtlasKV Output:
The opinion of MOROCCO is issue of synthetic biology should be considered as a new frontier.

KBLaM Output:
I'm not sure what you mean. Can you provide more context?

ICL Output:
The opinion of MOROCCO is synthetic biology should be considered as a new and emerging issue.

AtlasKV: What's Next?

(1) As we observe, this knowledge augmentation paradigm may good at dealing with background knowledge, which has lower requirements of accuracy. We may need to incorporate textual knowledge and parametric knowledge together to achieve a better knowledge system. (Like Meta just did.)

(2) Even though the GPU cost is lower, there could be more CPU and I/O cost. Maybe we can work on that to further improve the efficiency of this knowledge system.

(3) Knowledge can be aligned with LLMs. How about reasoning process? Can we align the reasoning process with LLMs?