Parallel Context-of-Experts Decoding for Retrieval Augmented Generation

ACL 2026 SAN DIEGO JULY 2–7

Why This Matters

RAG faces a core trade-off: richer cross-document reasoning versus system efficiency.

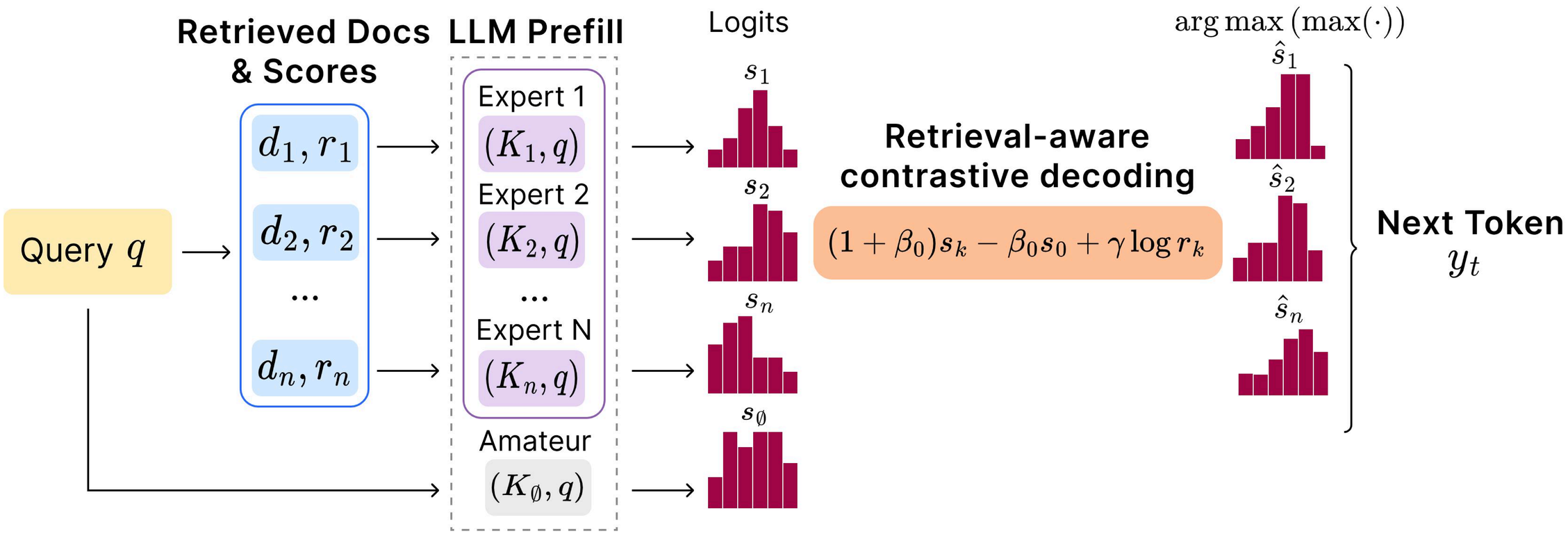
- Long prompts enable cross-document reasoning, but they are slow, and attention degrades with length.
- Independently encoded KV caches improve speed but break direct cross-document interaction.

Main Idea

Recover cross-document reasoning at decode time, while keeping documents independently cached.

- Encode each retrieved document independently as a reusable KV-cache expert.
- Route each token through the best-supported experts, enabling cross-document reasoning directly at decode time.

Do not concatenate contexts. Route every token through document experts.



Calibrate

Subtract the amateur prior to remove generic LM bias.

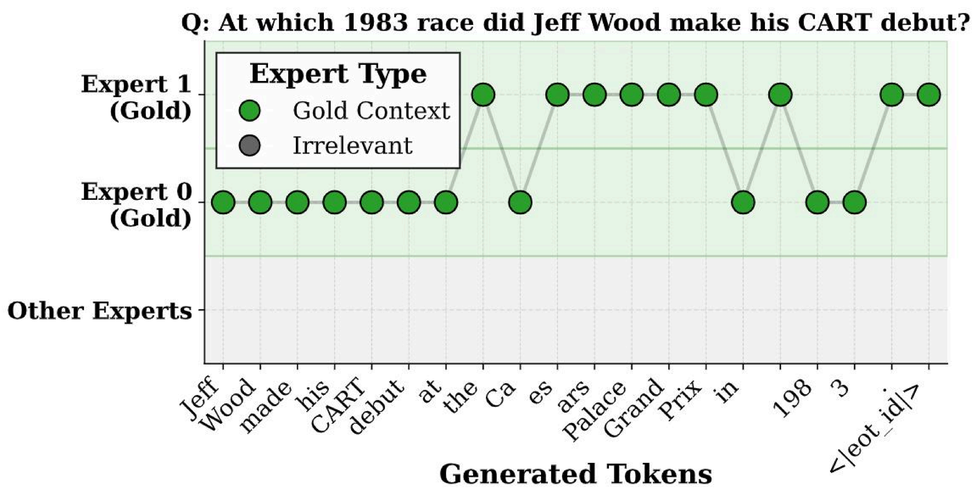
Gate

Add retrieval-score priors to favor better-supported document experts.

Route

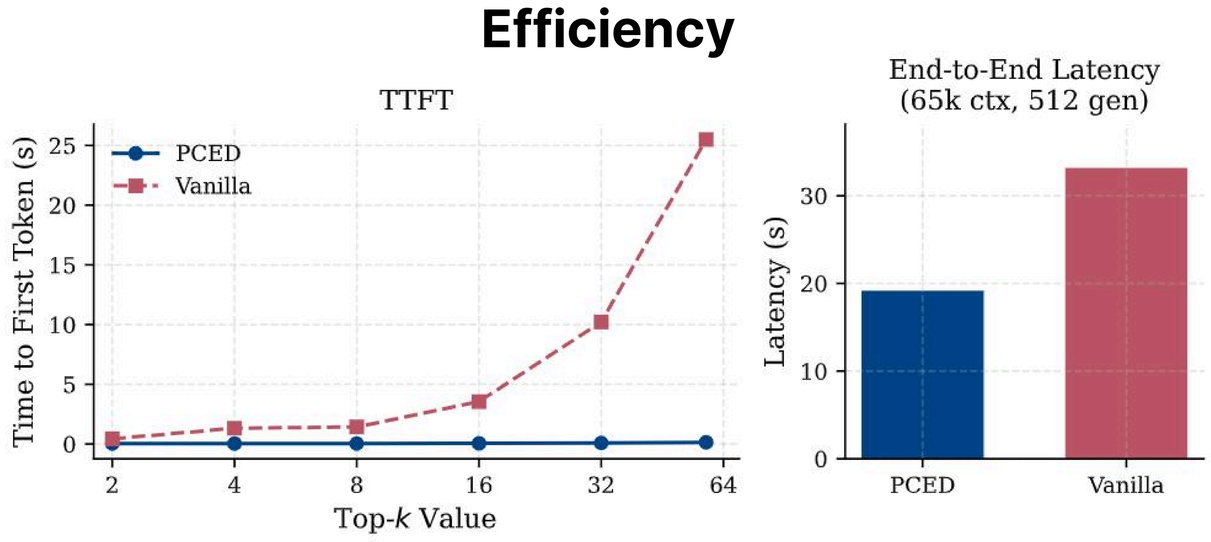
Select the expert with the strongest token-level support.

Decode-Time Reasoning



In **multi-doc QA**, PCED routes evidence token by token, and traces reveal when the model switches evidence sources.

Efficiency



0.14s vs 25.5s on TTFT.

1.7x better end-to-end latency.

Main Results

Model	Dataset	APE	MapRed.	PCED	Full Ctx
Llama-8B	HOTPOTQA	16	41	64	**73**
Llama-8B	NQ	9	51	**85**	82
Llama-8B	QAMPARI	7	68	77	**89**
Mistral-13B	HOTPOTQA	27	56	**66**	64
Mistral-13B	MUSIQUE	11	26	**36**	28
Mistral-13B	NQ	38	62	**81**	76

LOFT [1]: PCED outperforms APE [2] and MapRed. [3], sometimes surpassing full context.

Robustness to Distractors

Method	$k=4$	$k=8$	$k=16$	$k=32$	$k=64$
CLeHe	46	33	24	18	14
LeEns	48	36	27	19	15
PCED	**64**	**64**	**64**	**64**	**64**

Llama-3.1-8B on HotpotQA: soft expert mixtures, CLeHe and LeEns [4], degrade as k grows, while PCED remains stable.

Takeaways

Training-free | Retrieval-aware | Cache-native

References

[1] Lee et al. 2025, *LOFT: Scalable and More Realistic Long-Context Evaluation*, Findings NAACL 2025.
[2] Yang et al. 2025, *APE: Faster and Longer Context-Augmented Generation via Adaptive Parallel Encoding*, ICLR 2025.
[3] Zhou et al. 2025, *LLM×MapReduce: Simplified Long-Sequence Processing using Large Language Models*, ACL 2025.
[4] Qiu et al. 2025, *Entropy-Based Decoding for Retrieval-Augmented Large Language Models*, NAACL 2025.

SAP
EURECOM Sophia Antipolis

Affiliations
[1] SAP Labs, France
[2] EURECOM, France

Authors
Giulio Corallo [1, 2]
Paolo Papotti [2]

Contacts
giulio.corallo@sap.com
papotti@eurecom.fr