Hybrid-RACA: Hybrid Retrieval-Augmented Composition Assistance for Real-time Text Prediction

Menglin Xia*, Xuchao Zhang*, Camille Couturier, Guoqing Zheng, Saravan Rajmohan, Victor Rühle

Microsoft

{mollyxia, xuchaozhang} @ microsoft.com

What is Hybrid-RACA?

Hybrid-RACA is a system for **real-time text prediction** that efficiently combines a **cloud-based LLM and data** with a **small client-side model** through **retrieval augmented memory**.

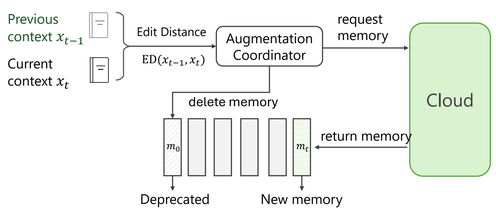
Key Features of Hybrid-RACA

- **Low latency** with hybrid modeling and asynchronous augmentation: Ensures fast response time by using the client model to make predictions
- **Enhanced utility** through hybrid retrieval augmentation and LLM-compressed memory: Improves the client model's suggestions
- **Minimized cloud-client communication:** Employs an augmentation coordinator to reduce the frequency of cloud-client interactions, leveraging LLM-compressed memory for enhanced efficiency.

Key Components of Hybrid-RACA

Step 1: The user types text on the client device.

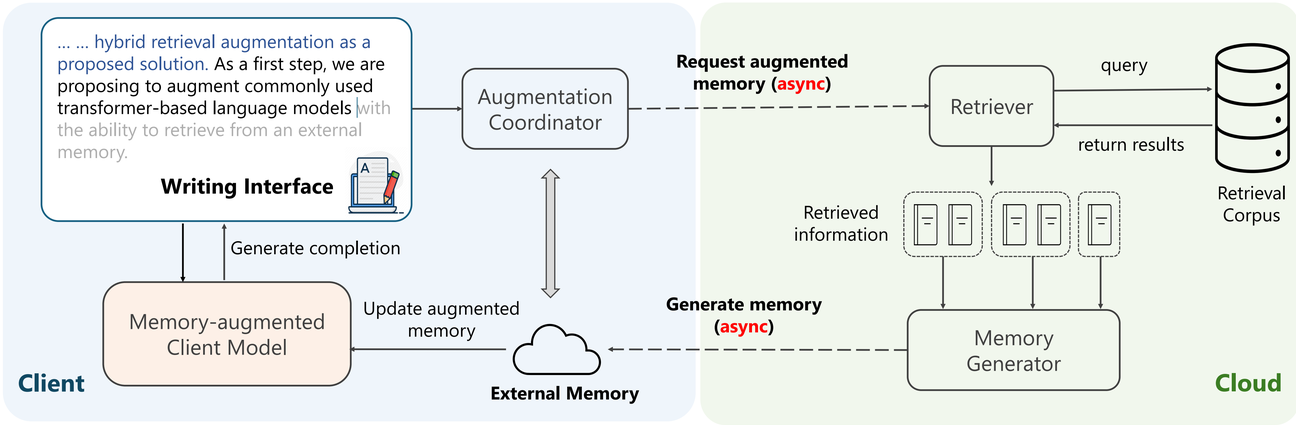
Step 2: The **Augmentation Coordinator** tracks changes to the writing context and sends an *asynchronous* request to the cloud if the changes exceed a predefined threshold.



Step 3: The **Retriever** fetches relevant documents and triggers the **Memory Generator**, a cloud-based LLM, to generate condensed *memory* that captures key takeaways from the retrieved content.

Step 4: The client-side memory is updated asynchronously with the new memory from the cloud.

*Step **: The **Memory-Augmented Client Model** makes more relevant predictions. It is instruction tuned to effectively leverage cloud –generated memory. *This step runs continuously and doesn't need to wait for step 2-4.*



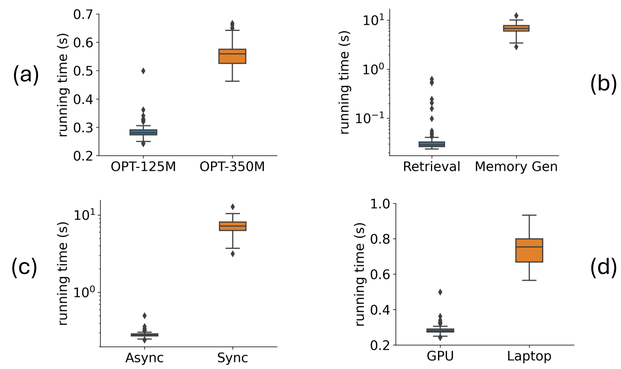
Utility Improvements

- **Substantial utility gains on in-domain evaluation**: Hybrid-RACA demonstrates strong text prediction performance on WikiText data with OPT-125M/OPT-350M
- **Consistent improvements on out-of-domain data** including Enron Emails, NIH ExPorter, Hacker News and Youtube Subtitles (results in paper)

		PPL	GLEU
OPT-125M	Vanilla OPT	9.3	11.4
	HybridRAG	4.3	12.8
	Hybrid-RACA w/o FT	3.8	14.7
	Hybrid-RACA FT	3.4	23.0
	Hybrid-RACA IT	**2.6**	**30.2**
OPT-350M	Vanilla OPT	7.4	13.2
	HybridRAG	3.6	15.4
	Hybrid-RACA w/o FT	3.3	17.6
	Hybrid-RACA FT	3.2	23.9
	Hybrid-RACA IT	**2.4**	**32.6**

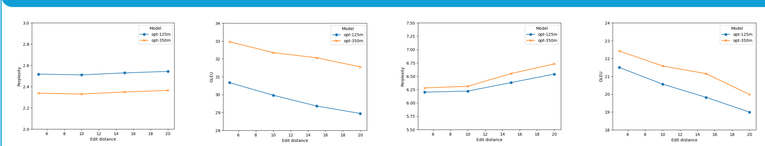
Model	GPT Score
GPT3.5	**7.73**
Vanilla OPT-125M	2.20
Vanilla OPT-350M	2.60
Hybrid-RACA IT OPT-125M	5.27
Hybrid-RACA IT OPT-350M	**5.49**

Latency Improvements



Highlights: Fig (c) showcases a remarkable **138x faster inference speed** achieved by our asynchronous design compared to a synchronous approach.

Trade-off with Asynchronous Memory Updates



- Performance decay with increased threshold for async memory update: As the memory becomes less fresh, performance gradually declines.
- Trade-off between freshness of memory and computation cost

About Us

We are the M365 Research Efficient AI team and AIOps team. We focus on advancing efficiency in AI systems and workflows. We're currently hiring PhD research interns for summer 2025 in both Cambridge, UK, and Redmond, US. Feel free to reach out!