ASSC 2026 · ASSOCIATION FOR THE SCIENTIFIC STUDY OF CONSCIOUSNESS

The Global Key-Value Workspace
A unifying consciousness theory for reasoning in LLMs

Zafeirios Fountas[1], Frederico Wieser[1], Martin Benfeghoul[1,2], Adnan Oomerjee[1,2], Jun Wang[2],

[1] Huawei Noah's Ark Lab, London · [2] AI Centre, Department of Computer Science, UCL

We build an explicit **global workspace** inside a pre-trained Large Language Model (LLM), via augmenting the Transformer architecture. We show it improves **reasoning** and increases the **synergy** between its components.

INTRODUCTION

A testbed for consciousness theories

- Consciousness theories are normally tested in brains, where clean intervention is hard.
- An **LLM** with an explicit workspace lets us build a theory's commitments in and manipulate them directly.
- We use it as a testbed for Global Workspace Theory, IIT, and how they relate.

RESULTS · PRETRAINING

Better, and cheaper

Over 7 standard reasoning tasks the workspace (BGT) beats the baselines on average, at lower validation loss and lower pretraining compute.

48.0%↑ avg accuracy · BGT (TF 46.8%)

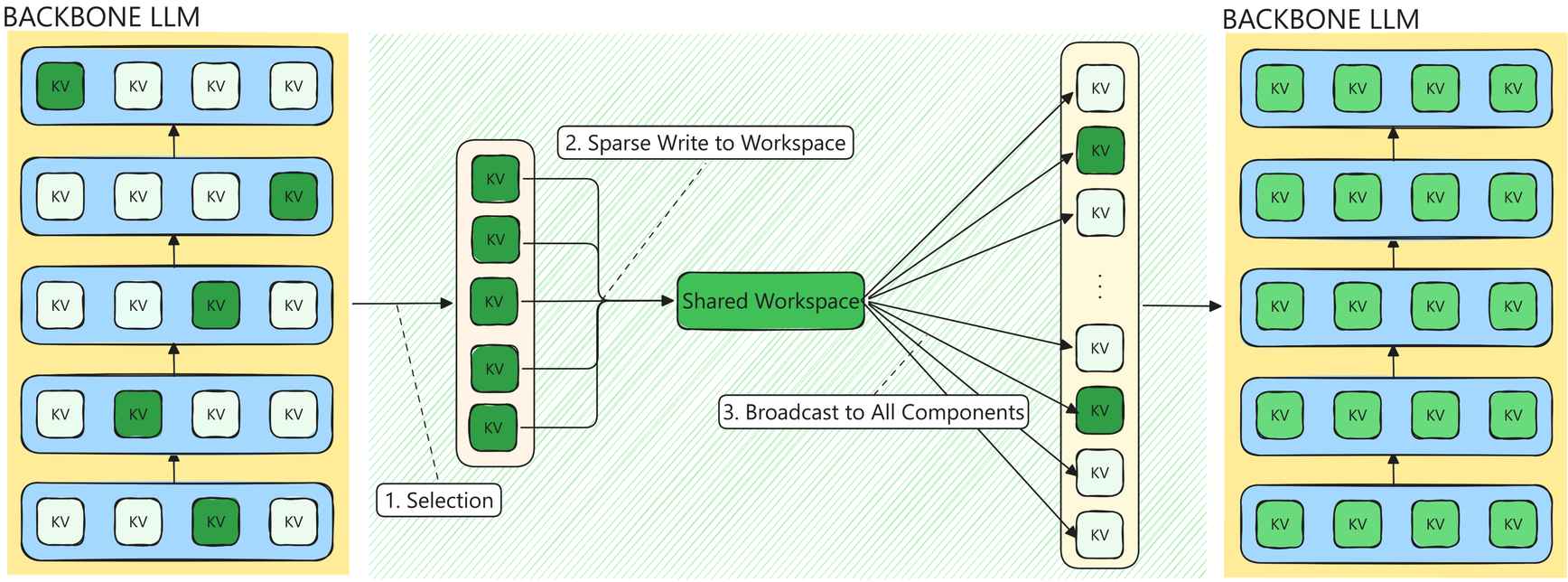
2.574 ↓ validation NLL · BGT (TF 2.603)

56.6 ↓ exaFLOPs · BGT (TF 58.5)

355–356M params · 20B tokens · matched across TF / BT / BGT.

ARCHITECTURE

Global Workspace Theory, made mechanistic



A capacity-limited spotlight selects salient components, writes them sparsely to a shared workspace, and broadcasts the result back to all components, iterated across layers. Unlike prior workspace networks trained from scratch, ours runs on a pretrained model's KV cache and we measure the integration it produces.

RESULTS · REASONING

The workspace improves math and logic reasoning

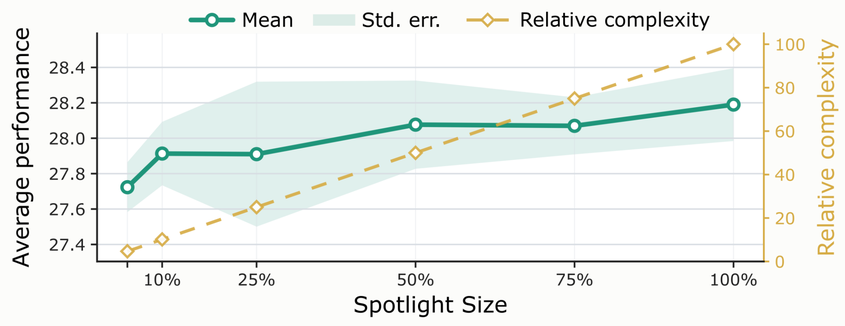
Method	GSM8K	SVAMP	GSM-Hd	LogiQA	Gaokao	AVG
Llama-3.2 1B						
SFT	33.00	42.00	8.00	28.90	23.90	**27.16**
BT	35.30	43.70	**8.20**	28.60	**25.60**	**28.28**
+GW	**35.60**	**46.70**	7.70	**29.50**	25.40	**28.98**
Llama-3.2 3B						
SFT	53.98	64.67	14.33	30.26	**31.05**	**38.86**
BT	57.09	**71.33**	14.63	30.26	30.20	**40.70**
+GW	**58.30**	69.67	**15.85**	**31.49**	30.48	**41.16**
Llama-3.1 8B						
SFT	13.12	23.33	3.11	**27.19**	**28.21**	**18.99**
BT	20.85	39.33	4.93	27.04	26.50	**23.73**
+GW	**21.76**	**40.00**	**5.38**	26.57	25.07	**23.76**
Qwen3-0.6B						
SFT	53.70	68.30	20.30	27.50	26.80	**39.32**
BT	54.80	68.70	**21.20**	27.20	26.50	**39.68**
+GW	**55.00**	**70.00**	21.10	**27.50**	**27.10**	**39.94**

TF transformer · BT Bottlenecked Transformer · +GW global workspace (dense broadcast). Average over the five tasks; same fine-tuning data throughout. Best per column shaded.

Across four backbones, broadcasting selected information improves multi-step reasoning. **The workspace does functional work, the access claim made concrete.**

RESULTS · PERFORMANCE VS COST

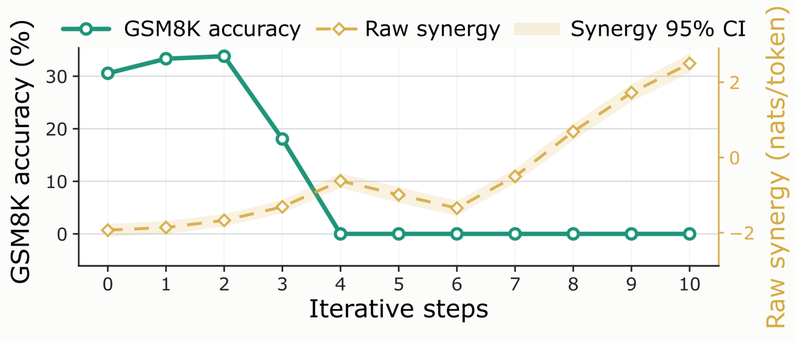
A narrow spotlight, at a fraction of the cost



We observe: Accuracy rises with width and is **highest at full broadcast**. A narrow spotlight recovers most of the accuracy at far lower compute.

RESULTS · PERFORMANCE VS INTEGRATION

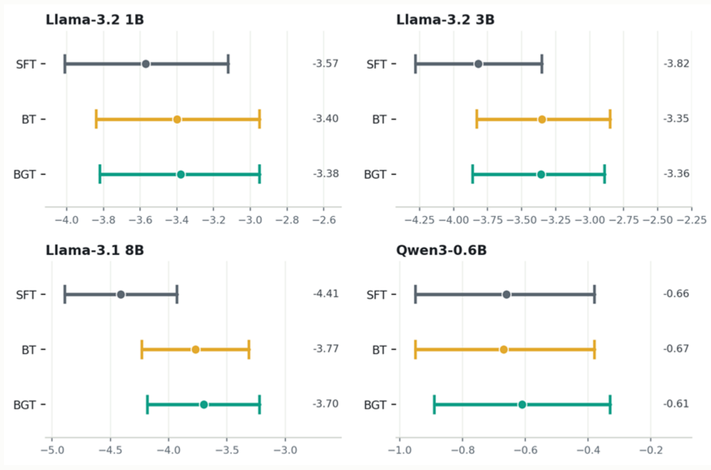
Integration rises to a peak, then falls



We observe: Synergy and performance have a complex relationship.

RESULTS · INTEGRATION

The workspace shifts components toward synergy

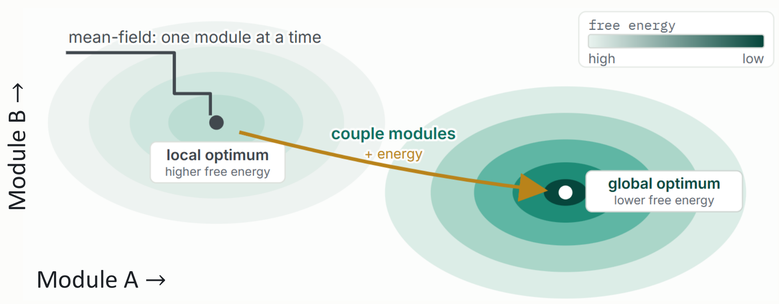


Synergy = information in the whole minus the sum of its parts, over balanced head partitions. We measure a synergy proxy across heads and compare across models (more positive = higher synergy). Results show that our model (BGT) tends towards higher synergy.

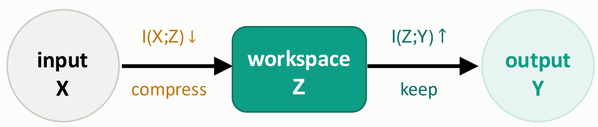
DISCUSSION

Why the workspace helps

Re-coupling. Specialised modules approximate a factorised, mean-field posterior that drops the dependencies between them. Synergy is the part no subset captures. The workspace re-couples a few to recover it and escape the local optima that factorised inference gets stuck in, at an energy cost.



Compression. Trained on the model's own loss, the workspace keeps what the latent tells us about the output while discarding input detail (data processing inequality). Compress the input, keep the output, the condition for generalisation.



Compression and synergy are orthogonal: how much of the input survives, versus how it is organised across modules. The workspace does both. We see synergy rise where broadcast helps; we have not yet shown it is the cause.

Open questions

- Are Global Workspace Theory and IIT actually rival accounts, or two views of one mechanism?
- Why is the workspace capacity-limited, and when should a bottleneck help beyond efficiency (distractors, interference, task-switching)?
- Does the workspace ignite, all-or-none, as GNWT predicts?
- What is the principled halting rule?
- Is the mean-field account of why workspaces help correct, trivial, or new?

We have the testbed. Help us design the experiment.

Register for the preprint and code Leave an email at the link and we will send the preprint, full results, and codebase when they are out.
Contact: zafeirios.fountas@huawei.com · frederico.wieser@huawei.com

Related Works Goyal & Bengio, Coordination Among Neural Modules Through a Shared Global Workspace, ICLR 2022 · VanRullen & Kanai 2021 · Luppi et al., synergistic workspace, 2024 · Safron, IWMT 2020 · Tishby, Pereira & Bialek, The Information Bottleneck Method, 1999 · Oomerjee et al., Bottlenecked Transformers, ICLR 2026 · Fountas et al., Predictive Forgetting, 2026