Not All Gap Correction Helps:

A Geometric View of External Information in LLM Inference

Dingzirui Wang · Xuanliang Zhang · Keyan Xu · Qingfu Zhu · Wanxiang Che
Harbin Institute of Technology

More corrective does not always mean more helpful.

Choose external information with the right direction and the right magnitude.

5.9% average relative error reduction

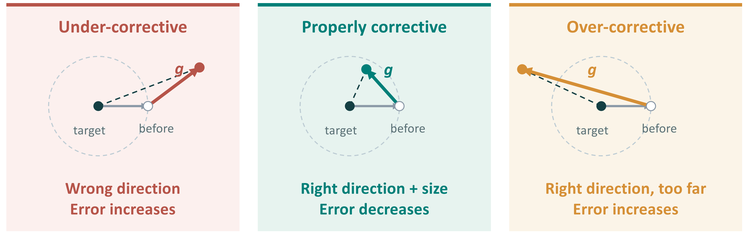
vs. main-table baselines · 26 settings

01 How context changes the answer

Demonstrations (ICL), retrieved knowledge (RAG), and memory (Mem) shift the model's output.

$$e(t, x) = e(x) + g(t, x)$$

x: query · t: context · e: baseline error · g: correction



Inside the circle = lower error. Colored dot = answer with context.

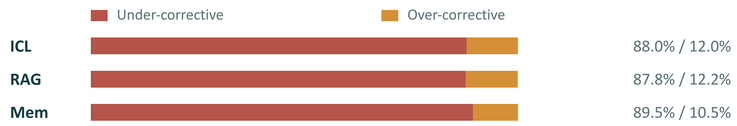
02 The direction–magnitude rule

$$\rho = \cos \angle(-e, g), \qquad r = \frac{\|g\|}{2\|e\|}$$

$$\|e + g\| < \|e\| \iff r < \rho$$

ρ measures alignment; r measures relative correction size.
For nonzero e, g; extends to multiple contexts and layers.

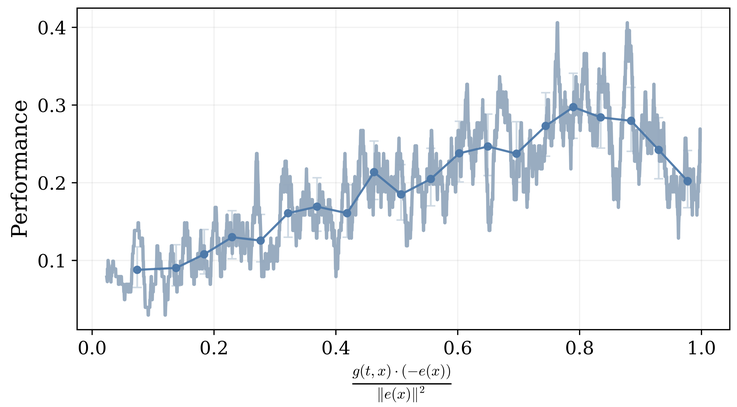
Among harmful cases, direction is the main failure.



Percentages are within the harmful-correction subset only.

03 Too much correction can hurt

Accuracy rises, then falls as corrective effect grows.

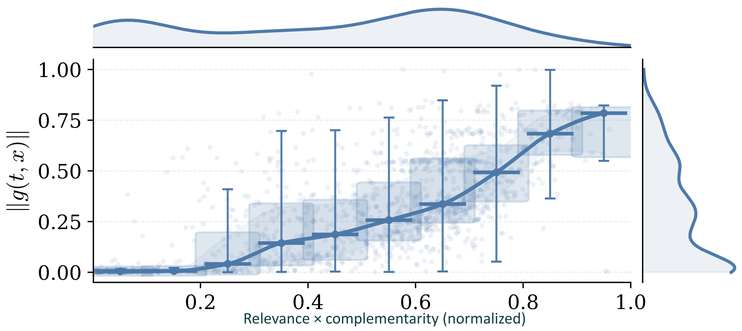


04 Attention is only half the story

Single-layer bound: attention × value-space complementarity.

$$\|g\| \le L_{\text{MLP}} \|W_O\| \, \alpha_{x \leftarrow t} \, \|v_t - v_x\|$$

Pearson r = 0.770



05 Select context in two stages

BM25 retrieves 32 candidates for each query.

1 Filter by predicted alignment

ELECTRA-Large uses the query, candidate, and the no-context answer to estimate ρ.

2 Rank by strength; select top-1

$$s(t, x) = \alpha_{x \leftarrow t} \, \|v_t - v_x\|$$

No ground-truth answer is needed at inference.

06 Gains across tasks and settings

Accuracy (%) · best main-table baseline → ours

Setting / dataset		Llama-3.1-8B	Qwen3-8B
ICL	MATH	47.6 → **49.6**	88.4 → **90.2**
	GPQA	43.0 → **46.5**	65.2 → **65.7**
	FinQA	51.6 → **53.1**	67.3 → **69.5**
RAG	CHAMP	24.8 → **27.4**	45.2 → **47.8**
	TheoremQA	40.0 → **42.5**	65.1 → **67.2**
	NQ	77.1 → **79.4**	75.6 → **78.9**
Mem	MATH	47.2 → **49.6**	88.2 → **89.0**
	CHAMP	15.2 → **18.1**	40.7 → **43.0**
	TheoremQA	27.7 → **30.4**	55.6 → **58.7**
	MMLU-Redux	64.9 → **67.1**	66.8 → **69.9**
	GPQA	32.3 → **34.8**	65.7 → **68.7**
	NQ	39.8 → **41.7**	45.3 → **47.5**
	FinQA	49.7 → **52.1**	61.5 → **63.4**

Ours: five-run results; uncertainty omitted here. Baselines are the best non-Ours scores in the main-paper tables.

Ablations: direction + strength beats either component alone.

Read the paper · Meet the author

OpenReview
Full paper & discussion

Homepage
Dingzirui Wang
Research & contact