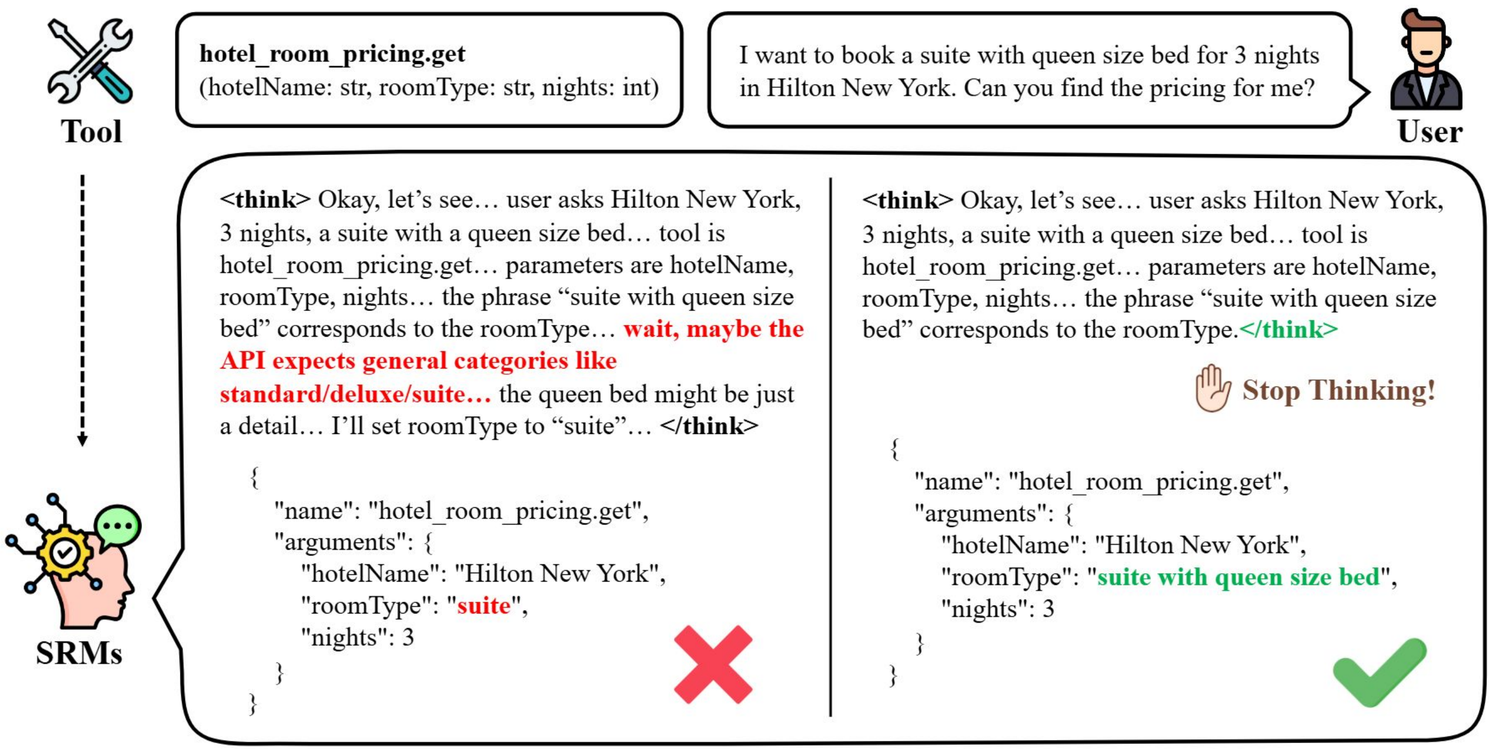
ThinkBrake - Mitigating Overthinking in Tool Reasoning

Minjae Oh*, Sangjun Song*, Seungkyu Lee*, Sungmin Jo, Yohan Jo
Graduate School of Data Science, Seoul National University

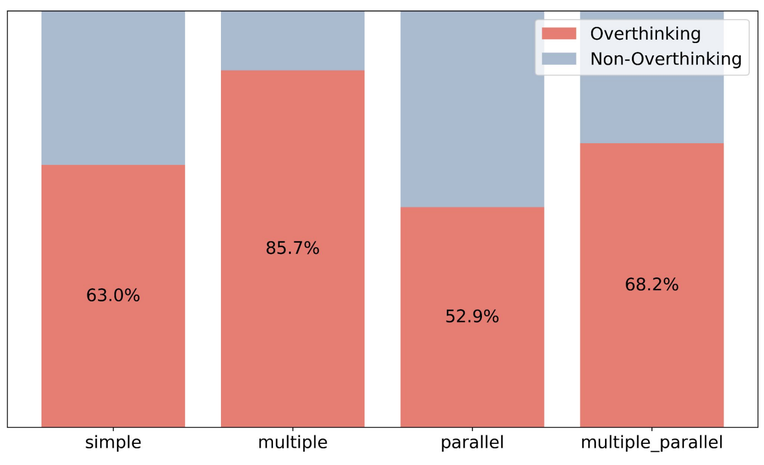
1. Motivation: Overthinking in Tool Reasoning



- Small Reasoning Models (SRM) generate a valid tool and argument configuration but subsequently override it by corrupting the arguments.
- Stopping at that moment by inserting a **</think>** token preserves the correct call.

Early-Stop Analysis

- Annotated the incorrect outputs to identify **overthinking** failures.
- A substantial fraction of failures stems from overthinking.



Oracle Setting

- For each reasoning trajectory, we inserted a **</think>** token **at the end of every sentence**, forcing the SRMs to stop thinking and produce an answer.

Method	simple Acc	simple ΔTok	multiple Acc	multiple ΔTok	parallel Acc	parallel ΔTok	multi-parallel Acc	multi-parallel ΔTok	Avg. Acc	Avg. ΔTok
Base	78.7	–	96.5	–	91.5	–	89.0	–	85.8	–
Oracle	91.5	-82.6%	98.5	-93.7%	96.0	-89.6%	95.5	-90.1%	94.2	-87.1%
Avg. Base Token Count	*1,154.5*		*985.0*		*1,199.9*		*1,621.2*		*1,214.1*	

3. Experiments

Mode	Method	simple Acc	simple ΔTok	multiple Acc	multiple ΔTok	parallel Acc	parallel ΔTok	multi-parallel Acc	multi-parallel ΔTok	Avg. Acc	Avg. ΔTok
Non-Live	Base	85.6	–	95.5	–	93.5	–	**90.5**	–	89.6	–
	NoWait	84.6	-56.6%	95.5	-61.5%	88.5	-44.4%	88.5	-53.5%	87.9	-54.8%
	ThinkLess	**88.7**	-100%	**98.0**	-100%	79.5	-100%	83.0	-100%	87.7	-100%
	`<tool_call>`	85.1	-15.7%	96.5	-25.5%	89.5	-39.6%	51.5	-52.5%	82.0	-28.0%
	THINKBRAKE (prob)	80.0	-25.5%	96.0	-51.3%	78.0	-26.9%	42.5	-33.4%	75.9	-31.2%
	THINKBRAKE	85.3	-25.6%	95.5	-17.5%	**95.5**	-26.8%	**90.5**	-28.2%	**89.8**	-24.9%
	Avg. Base Token Count	*1,154.5*		*985.0*		*1,199.9*		*1,621.2*		*1,214.1*	
Live	Base	86.4	–	**82.1**	–	**87.5**	–	79.2	–	**82.9**	–
	NoWait	**87.2**	-58.0%	81.1	-71.1%	81.3	-50.7%	75.0	-66.5%	82.2	-68.3%
	ThinkLess	77.1	-100%	76.4	-100%	37.5	-100%	45.8	-100%	75.5	-100%
	`<tool_call>`	86.4	-12.8%	81.9	-17.3%	75.0	-39.6%	62.5	-32.1%	82.3	-17.0%
	THINKBRAKE (prob)	81.4	-32.3%	76.9	-58.8%	62.5	-28.5%	29.2	-43.8%	76.8	-54.7%
	THINKBRAKE	85.7	-22.0%	81.4	-13.3%	**87.5**	-32.3%	**87.5**	-18.9%	82.4	-15.3%
	Avg. Base Token Count	*1,120.4*		*1,792.8*		*1,374.0*		*2,389.6*		*1,670.0*	

- Evaluate various heuristics using Qwen3-4B-Thinking-2507 on the Berkeley Function Calling Leaderboard (BFCL) non-live (v1) and live (v2) splits.
- Prior heuristics lead to **performance degradation**, especially in the parallel and multi-parallel categories, suggesting that useful reasoning is also pruned.
- Our method **maintains base accuracy or often improves it**, while **reducing tokens** by up to ~25% on non-live and up to ~15% on live splits.

2. ThinkBrake

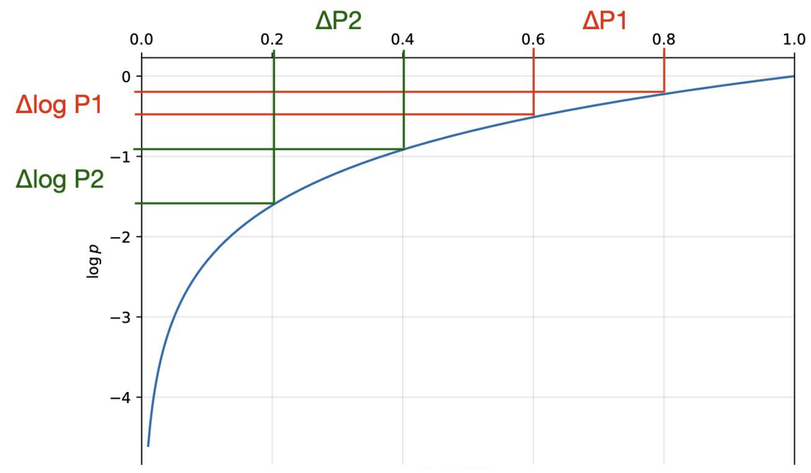
Log-Probability Margin

- Terminate thinking when the **log-probability margin** between the original token and the </think> token is small.

$$\log \frac{p_\theta(y_t^\star \mid x; y_{<t})}{p_\theta(y_{\texttt{</think>}} \mid x; y_{<t})} \leq \tau_{\text{threshold}}$$

Effects of Logarithm on ThinkBrake

- A meaningful gap only when both competing token probabilities are relatively high.
- Using a log-probability gap trigger activates only for ΔP1—where both the top token and </think> token have high probabilities.



4. Conclusions

- While reasoning leads to performance gains for SRMs, overthinking is a key source of errors.
- Terminating overthinking improves both token efficiency and performance.
- Overthinking is also prevalent in correct cases.

Limitations

- Evaluation on a single SRM family and one benchmark.
- Reliance on sentence boundary detection.

Future Work

- Test robustness across models/datasets.
- Not yet exhausted the upper performance bound and token reduction bound suggested by the oracle setting, leaving space for further improved pruning methods.

References
- Li G. et al. *ThinkLess: Reducing Reasoning Redundancy without Training.* arXiv, 2025.
- Patil S. G. et al. *The Berkeley Function Calling Leaderboard (BFCL).* ICML, 2025.
- Wang C. et al. *Wait, We Don't Need to "Wait"! Removing Thinking Tokens.* arXiv, 2025.
- Zhang X. et al. *Making Small LMs Efficient Reasoners via Intervention, Supervision, RL.* arXiv, 2025.

NEURAL INFORMATION PROCESSING SYSTEMS