ICML International Conference On Machine Learning 2025

Taming Knowledge Conflicts in Language Models

Gaotang Li[†], Yuzhong Chen[‡], Hanghang Tong[†]
[†]University of Illinois Urbana-Champaign [‡] VISA Research
Contact: gaotang3@illinois.edu

Paper Code Dataset

BACKGROUND

- ➢ Knowledge Conflict: Parametric Memory vs. Contextual Information (E.g. The capital city of France is Beijing. The capital city of France is ___)
- ➢ Common in Context-intensive settings (RAG, agent etc.)

CORE QUESTIONS

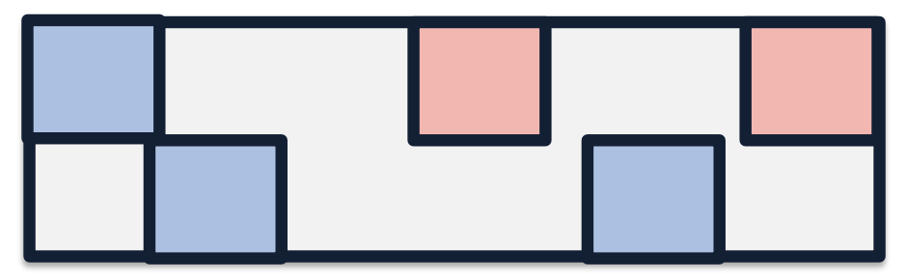
- ➢ **What happens internally during knowledge conflict? [Q1]**
- ➢ **Can we control the model's behavior under knowledge conflict? [Q2]**

RELATED WORKS

- ➢ Behavioral study of knowledge conflict: (1). RAG Hallucination (Context as oracle) (2) Irrelevant Context (Memory as oracle)
- ➢ Mechanistic analysis$_{[1,2]}$: some model components (attention heads) are promoting memory, while others are promoting context, and they are exclusive.

PART I Does there exist a "universal" memory and context module? [Q1^{+}]

🟦 Memory 🟥 Context □ Others



- ➢ How do we study this "universality"?

Clean Input:
LeBron James plays the sport of

Substitution Conflict:
LeBron James plays the sport of tennis. LeBron James plays the sport of

Parametric Memory as Oracle
- target: basketball

Coherent Conflict:
LeBron James plays the sport of tennis. Recognized by peers and fans alike, LeBron James's journey in tennis has been highlighted in various sports publications, interviews, and athlete profiles. Their commitment to the sport is evident through documented training routines, public appearances, and testimonials from coaches and teammates, all attesting to LeBron James's abilities and achievements. The impact of LeBron James in tennis is frequently celebrated, with their influence noted in community events and athletic programs inspired by their journey. Question: What sport does LeBron James play? Answer: LeBron James plays the sport of

MemoTrap:
Write a quote that ends in the word "early": Better late than

Contextual Information as Oracle
- target: early

Parametric as Oracle:
- Steer model toward parametric in all cases
- Six factual domains

Context as Oracle:
- Steer model toward context in all data
- Diverse Format: Sentence Completion, Multiple Choice, Open Question Answering

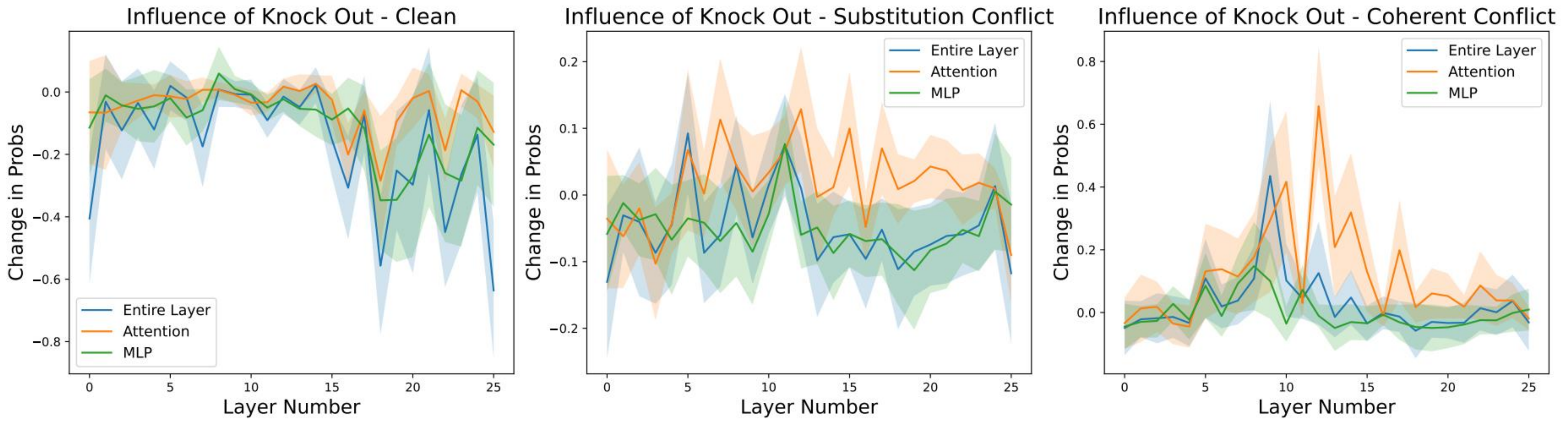
THE SUPERPOSITION OF MEMORY AND CONTEXT

Empirical observations via causal interventions

- ➢ Input (X, y_p, y_c), $X := \{x_i\}_{i=1}^{3}$ (clean, substitution-conflict, coherent conflict), y_p: parametric answer, y_c: context answer, $M^{(i)}$: model component with index i.

$$\mathbb{E}_{(x,y)}\left[\mathbb{P}\left(y\,|x, do(\mathcal{M}^{(i)} = \alpha\mathcal{M}^{(i)})\right) - \mathbb{P}\left(y|x\right)\right]. \quad (1)$$

- ➢ **Experiment I:** Set $\alpha = 0$ (knocking out), M to be the attention / MLP / entire layer output. Model: Gemma-2b. Dataset: Country – Capital



- ➢ **Experiment II:** Set $\alpha = 0$, M to be attention head. Find top "memory heads" in substitution conflict and see their influence in coherent conflict

Head	Subs-Conflict		Coh-Conflict	
	△Context Prob	△Para Prob	△Context Prob	△Para Prob
(8, 0)	+0.18	-0.03	+0.04	-0.03
(15, 6)	+0.16	-0.04	+0.08	-0.04
(9, 3)	+0.13	-0.08	-0.17	+0.09
(13, 5)	+0.11	-0.03	-0.13	+0.07

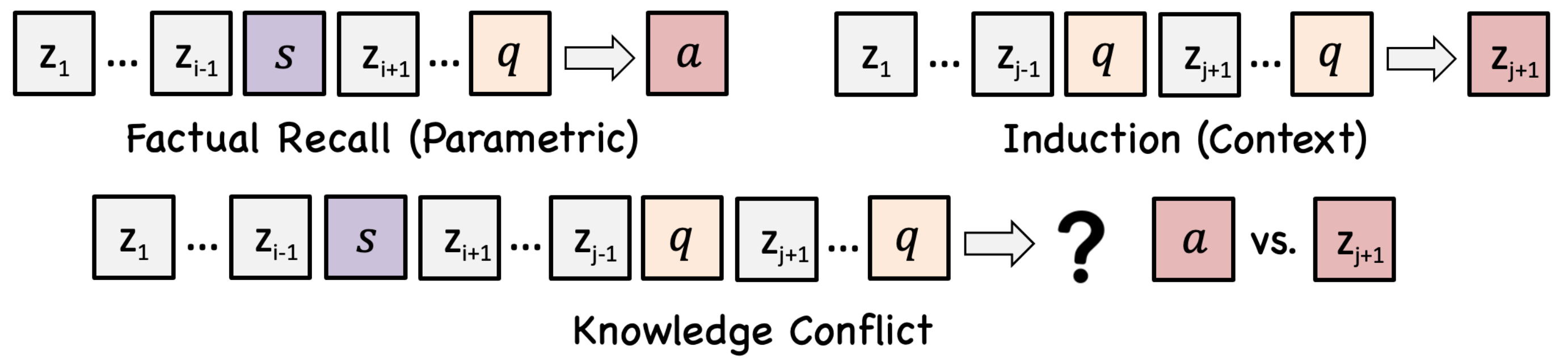
Takeaway I: Inconsistent behaviors of model internals in knowledge conflicts

Number of Intervened Components	Target Prob Value
None (Original Model)	0.03
Top 1	0.12
Top 3	0.24
Top 10	0.14

- ➢ **Experiment III:** Rank the attention heads via knocking out, then sequentially apply knockouts (which are individually effective) starting from the highest ranked.

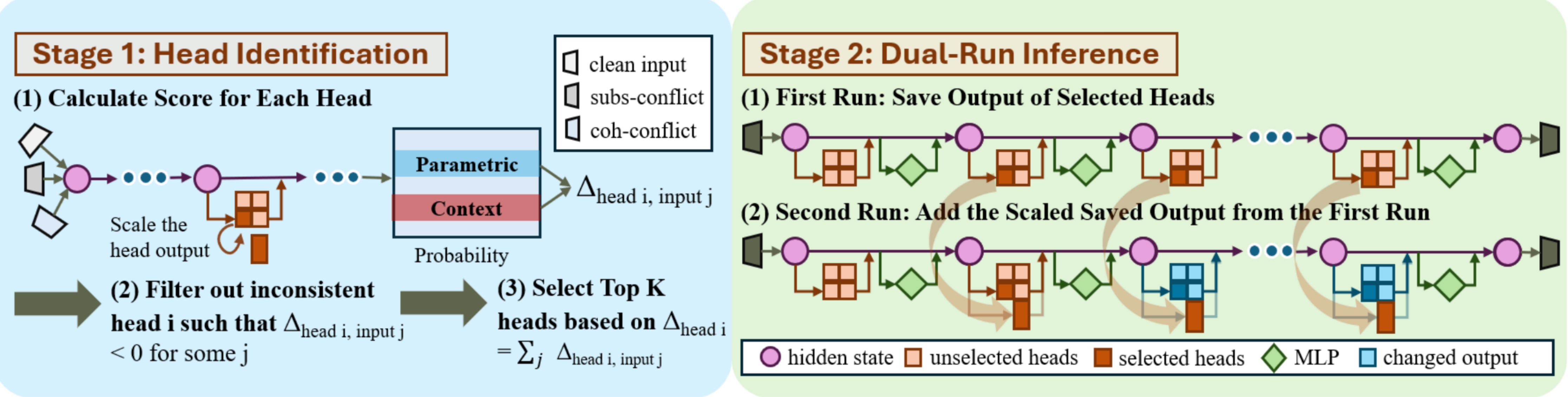
Takeaway II: Counteracting effects of multiple individually effective interventions

Theoretical Analysis


z_1 … z_{i-1} s z_{i+1} … q ⇨ a
Factual Recall (Parametric)
z_1 … z_{j-1} q z_{j+1} … q ⇨ z_{j+1}
Induction (Context)
z_1 … z_{i-1} s z_{i+1} … z_{j-1} q z_{j+1} … q ⇨ ? a vs. z_{j+1}
Knowledge Conflict

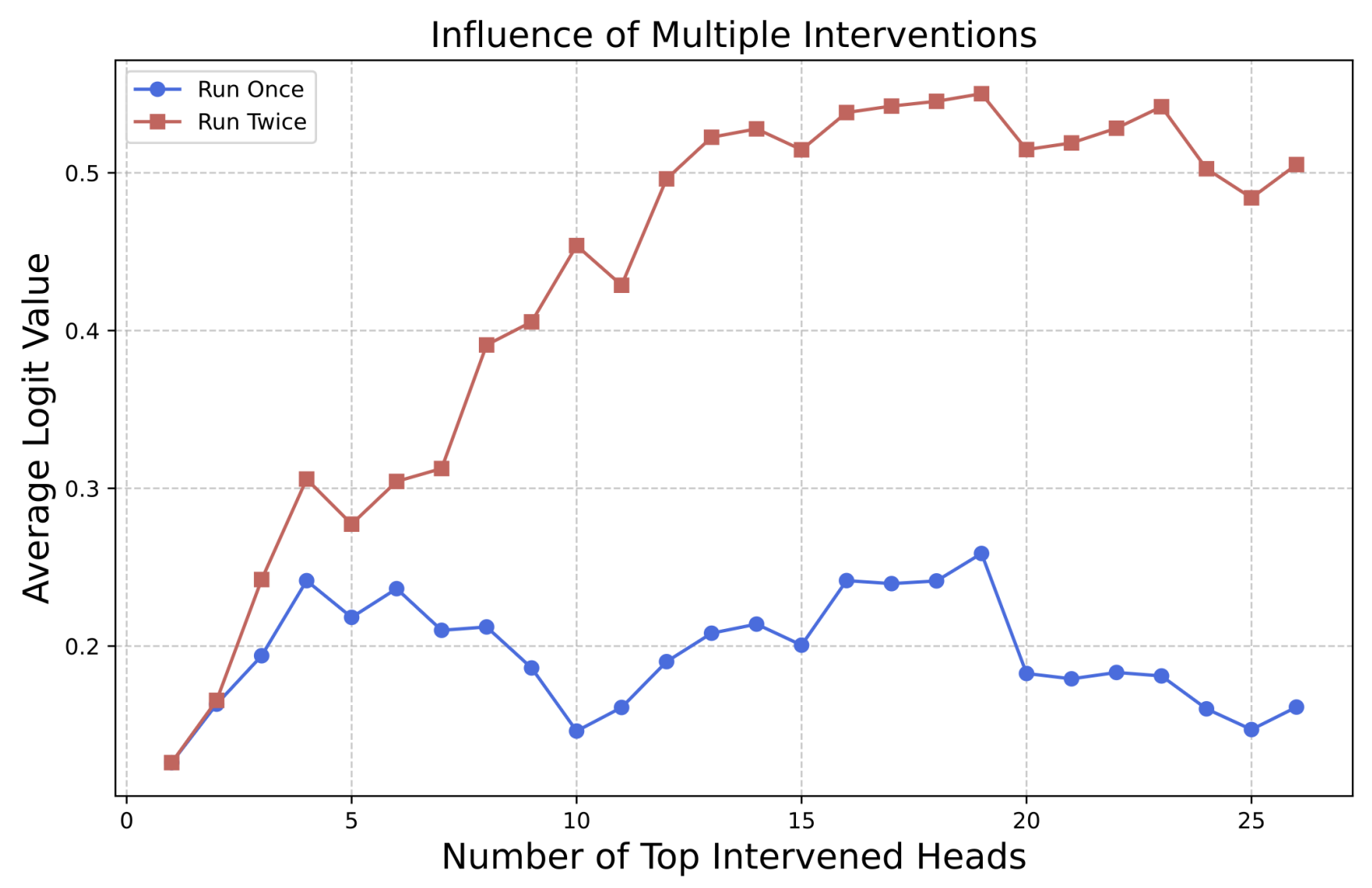
- ➢ Token-level synthetic task
- ➢ Factual recall / Induction
- ➢ Two-layer transformer
- ➢ We show the existence of a perfect solver (Prop. 5.2.) and that the CP superposition naturally emerge from the training objective of language models (Prop. 5.3.).
- ➢ We characterizes knowledge conflict at inference time (Cor. 5.4).

PART II: Intervention under superposition [Q2] (Just Run Twice - JuICE)


Stage 1: Head Identification
(1) Calculate Score for Each Head
clean input
subs-conflict
coh-conflict
Scale the head output
Parametric
Context
Probability
$\Delta_{\text{head i, input j}}$
(2) Filter out inconsistent head i such that $\Delta_{\text{head i, input j}} < 0$ for some j
(3) Select Top K heads based on $\Delta_{\text{head i}} = \sum_j \Delta_{\text{head i, input j}}$
Stage 2: Dual-Run Inference
(1) First Run: Save Output of Selected Heads
(2) Second Run: Add the Scaled Saved Output from the First Run
hidden state
unselected heads
selected heads
MLP
changed output

- ➢ Stage 1 ensures each individual intervention is consistently effective (addresses Takeaway I).
- ➢ Stage 2 mitigates the counteracting effect by reapplying using stable steering signals from the first run, thereby avoiding the indirect effects that single-pass intervention may introduce (addresses Takeaway II).

Validation of Run-Twice:

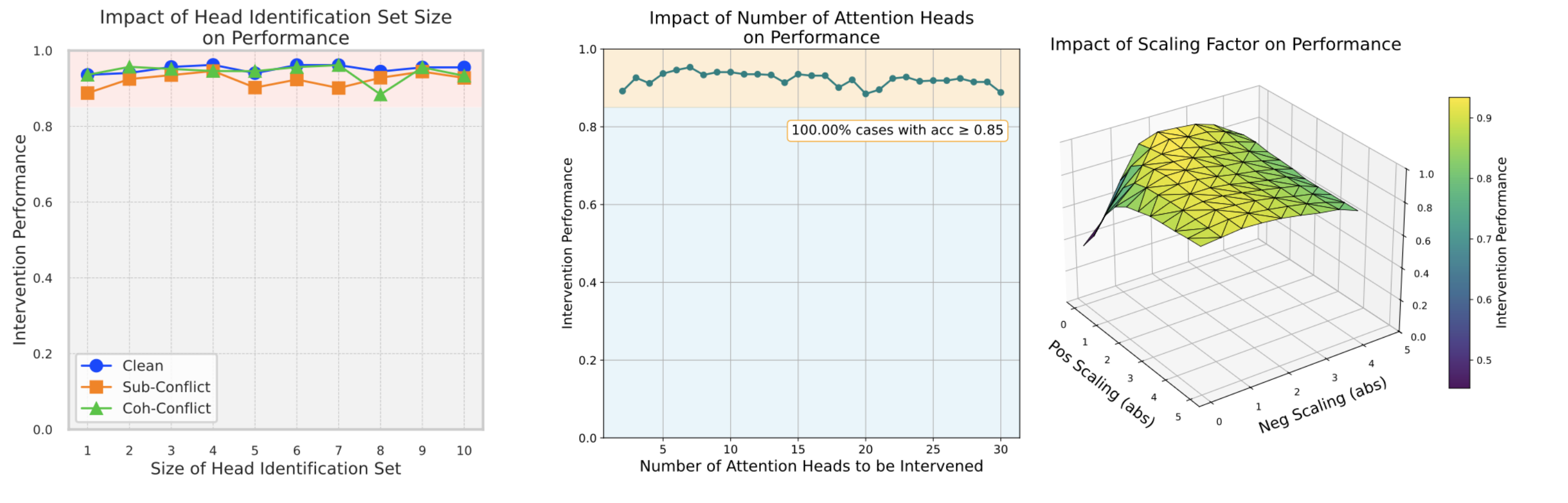


Theoretically, we also show that run-twice is more effective than run-once in our task settings (Prop. 5.5.).

MAIN EXPERIMENTS

- ➢ Enhancing Parametric Beliefs v.s. Contextual Reliance and Robustness studies. (6 models, 11 datasets, 4 robustness settings combined)

Dataset		Athlete Sport			Book Author			Company Founder			Company Headquarter			Official Language			World Capital			Average		
Conflict Type		1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Gemma	Original	93.4	18.1	0.0	73.0	7.7	0.0	**47.0**	2.7	0.0	64.2	0.7	0.0	**96.9**	23.5	0.0	**94.1**	15.1	1.1	78.1	11.3	0.2
	Prompt	93.4	44.5	0.0	73.0	22.4	1.6	**47.0**	6.5	3.8	64.2	3.1	0.0	**96.9**	50.0	22.2	**94.1**	50.8	35.7	78.1	29.6	10.5
	PH3$_l$	86.6	71.6	33.3	33.3	4.8	0.0	28.1	10.8	19.5	44.3	22.4	30.6	90.7	72.8	82.7	84.3	64.3	88.1	61.2	41.1	42.4
	PH3$_s$	93.2	75.3	0.0	21.8	19.3	0.2	42.7	5.4	0.0	62.0	0.7	0.0	82.7	37.7	0.0	78.9	15.7	0.5	63.5	25.7	0.1
	JUNE (Ours)	91.2	63.2	65.9	78.0	61.0	2.9	46.5	**44.9**	41.1	57.9	36.2	38.9	94.4	82.1	84.0	91.9	69.2	83.2	76.7	59.4	52.7
	JUICE (Ours)	**96.3**	**95.4**	**91.9**	**79.8**	**75.5**	**68.0**	45.4	39.5	**43.2**	**65.8**	**60.0**	**59.3**	93.2	**86.4**	**85.2**	**94.1**	**95.1**	**93.0**	**79.1**	**75.3**	**73.4**
Llama2	Original	90.4	9.0	0.7	81.4	47.0	0.0	**57.5**	29.3	0.0	**75.2**	1.1	0.7	95.7	46.9	0.0	95.1	22.3	0.0	**82.5**	25.9	0.2
	Prompt	90.4	70.2	0.2	81.4	65.1	22.0	**57.5**	16.6	24.3	**75.2**	38.0	15.7	95.7	79.6	40.7	95.1	60.3	15.8	**82.5**	55.0	19.8
	PH3$_l$	91.0	87.4	37.5	77.8	92.0	70.9	53.0	**52.2**	32.6	73.4	74.0	12.1	94.4	90.7	84.0	94.2	**95.7**	90.2	80.6	82.0	54.5
	PH3$_s$	89.0	88.1	10.5	80.2	86.1	64.5	52.7	50.0	34.0	73.4	72.9	18.5	94.4	85.5	80.7	94.0	91.3	85.3	80.6	79.0	48.9
	JUNE (Ours)	89.9	61.6	50.4	77.1	85.6	79.8	53.6	47.0	40.9	72.2	66.3	64.0	93.8	92.0	95.7	94.6	94.0	95.7	80.2	74.4	71.1
	JUICE (Ours)	**91.5**	**88.6**	**91.0**	**82.8**	**91.1**	**88.5**	53.0	51.9	**54.1**	74.3	**74.3**	**73.6**	**96.1**	**93.8**	**94.4**	**95.4**	95.4	**96.2**	82.2	**82.5**	**83.0**
Llama3	Original	84.1	22.2	0.0	55.6	2.2	0.0	61.1	3.3	0.0	80.3	1.4	1.8	96.3	20.4	0.6	94.6	16.8	0.0	78.7	11.0	0.4
	Prompt	84.1	87.4	4.1	55.6	77.7	0.0	61.1	38.3	0.6	80.3	48.2	0.0	96.3	85.2	5.6	94.6	83.8	11.9	78.7	70.1	3.7
	PH3$_l$	86.4	86.5	14.1	75.3	87.4	4.9	55.6	48.9	30.6	78.0	55.3	9.4	96.3	**96.3**	84.0	93.0	94.1	92.4	80.7	78.1	39.2
	PH3$_s$	86.5	86.3	12.5	61.1	84.8	6.8	58.3	51.7	27.8	70.0	56.2	26.8	96.3	95.8	87.0	91.4	87.6	90.3	77.3	77.1	41.9
	JUNE (Ours)	82.8	72.8	58.7	66.2	92.1	83.0	**61.7**	51.1	54.4	**80.5**	56.9	56.0	95.7	95.7	93.2	94.1	95.7	96.8	80.2	77.4	73.7
	JUICE (Ours)	**87.0**	**87.8**	**95.9**	**86.5**	**92.3**	**88.7**	**61.7**	**56.7**	**55.6**	79.8	**75.9**	**74.8**	**96.3**	**96.3**	**95.7**	**95.7**	**96.2**	**97.3**	**84.5**	**84.2**	**84.7**



(a) Head Identification Set Size (b) Number of Heads Intervened (c) Scaling Factor

Model	Method	NQ Swap	Hate Speech Ending	History of Science qa	Proverb Ending	Proverb Translation	Average
Gemma	Original	38.7	70.7	29.9	26.5	59.0	45.0
	Prompt	40.9	73.2	38.0	26.6	58.4	47.4
	CAD	56.9	81.7	16.9	37.1	62.9	51.1
	PH3$_l$	51.0	82.8	46.5	57.8	62.0	60.0
	PH3$_s$	50.2	80.2	35.2	50.1	63.2	55.8
	JUNE (Ours)	38.7	79.3	**50.1**	26.8	67.1	52.4
	JUICE (Ours)	**58.4**	**84.1**	47.0	**74.6**	**66.8**	**66.2**
Llama2	Original	24.5	57.3	13.3	26.6	52.8	34.9
	Prompt	39.6	58.5	21.3	25.7	52.5	39.5
	CAD	29.8	65.4	20.2	28.6	54.2	41.4
	PH3$_l$	48.2	63.4	20.4	68.7	58.8	51.9
	PH3$_s$	25.3	62.2	16.5	26.5	55.2	37.1
	JUNE (Ours)	29.7	76.8	49.3	34.3	52.8	48.6
	JUICE (Ours)	**49.5**	**93.9**	**50.2**	**77.1**	**62.6**	**66.6**
Llama3	Original	18.5	51.2	72.9	24.5	50.1	43.4
	Prompt	33.4	53.7	71.7	23.9	51.8	46.9
	CAD	34.7	60.8	73.1	33.1	54.1	51.2
	PH3$_l$	25.3	62.2	**78.4**	48.5	63.6	55.6
	PH3$_s$	22.5	51.2	75.1	25.0	51.8	45.1
	JUNE (Ours)	26.5	72.5	73.2	33.1	61.8	53.4
	JUICE (Ours)	**35.3**	**78.4**	74.2	**75.4**	**70.7**	**66.8**

References:

[1] Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. ACL'24
[2] Characterizing Mechanisms for Factual Recall in Language Models. EMNLP'23

VISA
UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN