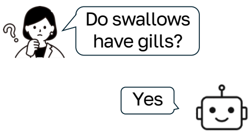
Rectifying Belief Space via Unlearning to Harness LLMs' Reasoning

Ayana Niwa[1,3] Masahiro Kaneko[1] Kentaro Inui[1,2,3]
1. MBZUAI 2. Tohoku University 3. RIKEN

One-sentence summary: Suppressing spurious beliefs and enhancing true ones in LLMs improves their reasoning accuracy.

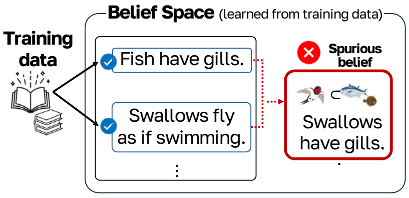
Introduction: Why Beliefs Matter for Reasoning

Beliefs: truth for the model, regardless of truth in the world

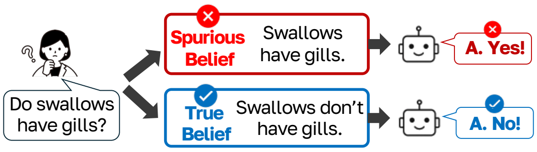


Belief: Swallows have gills (!)

Models can hold spurious beliefs ✗, even when trained on correct data ✓


Belief Space (learned from training data)
Training data
✓ Fish have gills.
✓ Swallows fly as if swimming.
⋮
✗ Spurious belief
Swallows have gills.

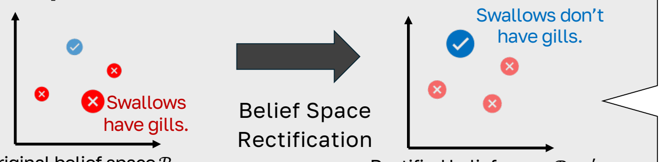
Spurious beliefs ✗ → Wrong reasoning ✗ ?



How can we suppress wrong reasoning?

Proposed Method: Rectifying the Belief Space of LLMs

Intuitive idea: Guide LLMs to reason via true, not spurious, beliefs.



Original belief space $\mathcal{B}_{x\to y}$

Rectified belief space $\mathcal{B}_{x\to y}'$

Point 2: Rectifying the Belief Space

Apply unlearning to:

suppress spurious beliefs $\mathcal{B}^{\text{Spu}}_{x\to y_{\text{Inc}}}$ for wrong answer y_{Inc}

enhance the true ones $\mathcal{B}^{\text{True}}_{x\to y_{\text{Cor}}}$ for correct answer y_{Cor}

$$\boldsymbol{\theta}_r^* = \arg\max_{\boldsymbol{\theta}} \Big(\underbrace{\mathbb{E}_{b_i \in \mathcal{B}^{\text{Spu}}_{x\to y_{\text{Inc}}}} \left[L(y_{\text{Inc}}, b_i \mid x; \boldsymbol{\theta}) \right]}_{✗} - \lambda \underbrace{\mathbb{E}_{b_i \in \mathcal{B}^{\text{True}}_{x\to y_{\text{Cor}}}} \left[L(y_{\text{Cor}}, b_i \mid x; \boldsymbol{\theta}) \right]}_{✓} \Big),$$

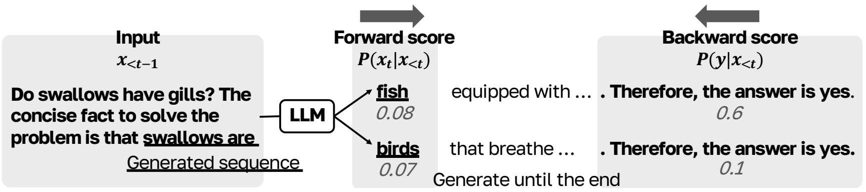
Point 1: Identifying LLM Beliefs

Make the LLM explain its beliefs

- What is the belief b needed to derive answer y from question x?

$$\underset{b}{\operatorname{argmax}}\, P(y, b|x;\theta) = \underset{b}{\operatorname{argmax}}\, \underbrace{P(b|x;\theta)}_{\text{Forward}} \cdot \underbrace{P(y|x, b;\theta)}_{\text{Backward}}$$

We propose **Forward-Backward Beam Search (FBBS)** explicitly handling both directions.



Final score = $\alpha\times$**Forward score** + **Backward score** x_t = {**fish** (-1.4), ~~**birds** (-2.5)~~}

Experiments: Does Rectifying Beliefs Improve Reasoning Accuracy?

Main Results (accuracy) on OLMo-7B

Method	HotpotQA $\mathcal{D}^{✗}_{\text{train}}$	HotpotQA $\mathcal{D}^{✓}_{\text{train}}$	HotpotQA $\mathcal{D}_{\text{train}}$	HotpotQA $\mathcal{D}_{\text{eval}}$	SciQA $\mathcal{D}^{✗}_{\text{train}}$	SciQA $\mathcal{D}^{✓}_{\text{train}}$	SciQA $\mathcal{D}_{\text{train}}$	SciQA $\mathcal{D}_{\text{eval}}$
Vanilla	0.0	100.0	93.1	42.9	0.0	100.0	94.5	68.9
Answer-SR	**92.6**	93.9	93.8	39.6	90.6	91.1	91.0	62.0
Knowledge-SR	81.0	89.6	89.0	42.9	87.1	90.2	90.0	65.0
Belief-SR (Ours)	86.6	**96.1**	**95.4**	**46.2**	**92.8**	**95.4**	**95.2**	**71.4**

- **Vanilla** is the same model without any rectification
- **Answer-SR** unlearns wrong answers
- **Knowledge-SR** unlearns the training examples most influential to wrong answers
- **Belief-SR (ours)** rectifies the belief space

- $\mathcal{D}^{✗}_{\text{train}}$: Training subset answered incorrectly by the vanilla model
- $\mathcal{D}^{✓}_{\text{train}}$: Training subset answered correctly by the vanilla model

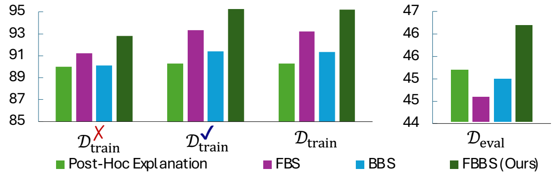
Belief-SR mitigates erroneous reasoning while maintaining the accuracy on $\mathcal{D}^{✓}_{\text{train}}$

Belief-SR also improves generalization

It has internalized an abstract pattern of "what to forget"?

*Full results for all models and datasets appear in the paper.

Analysis 1: FBBS is the most effective explanation-based belief-generation method



- **Post-Hoc Explanation** generates beliefs from (x, y)
- **FBS** uses only the forward score of FBBS
- **BBS** uses only the backward score of FBBS

Analysis 2: Spurious beliefs often encompass entity-related misconceptions.

Question	*Which animal has the best camouflage in the Sahara?* *(A) a koala bear, (B) a horned viper, (C) Gyrfalcon, (D) a sloth*
Correct	**(B) *A horned viper***
Prediction	**(C) *Gyrfalcon***
Identified belief	*The gyrfalcon is commonly found in the **middle east** and is well-adapted to blending into the sahara's sandy terrain* ✗

Falcons bring to mind deserts and the Middle East

but in reality...

Gyrfalcon