FY2025/2025年度

Natural Language Understanding Team

Kentaro Inui

自然言語理解チーム　乾 健太郎

RIKEN

Center for Advanced Intelligence Project

Mechanistic Understanding of GenAI

Vision and Social Impact

The adoption of GenAI has opened many new challenges. Mechanistic understanding of GenAI is needed to ensure safety, fairness, and trustworthiness.
Our vision is to achieve this understanding by analyzing models at three levels of granularity:

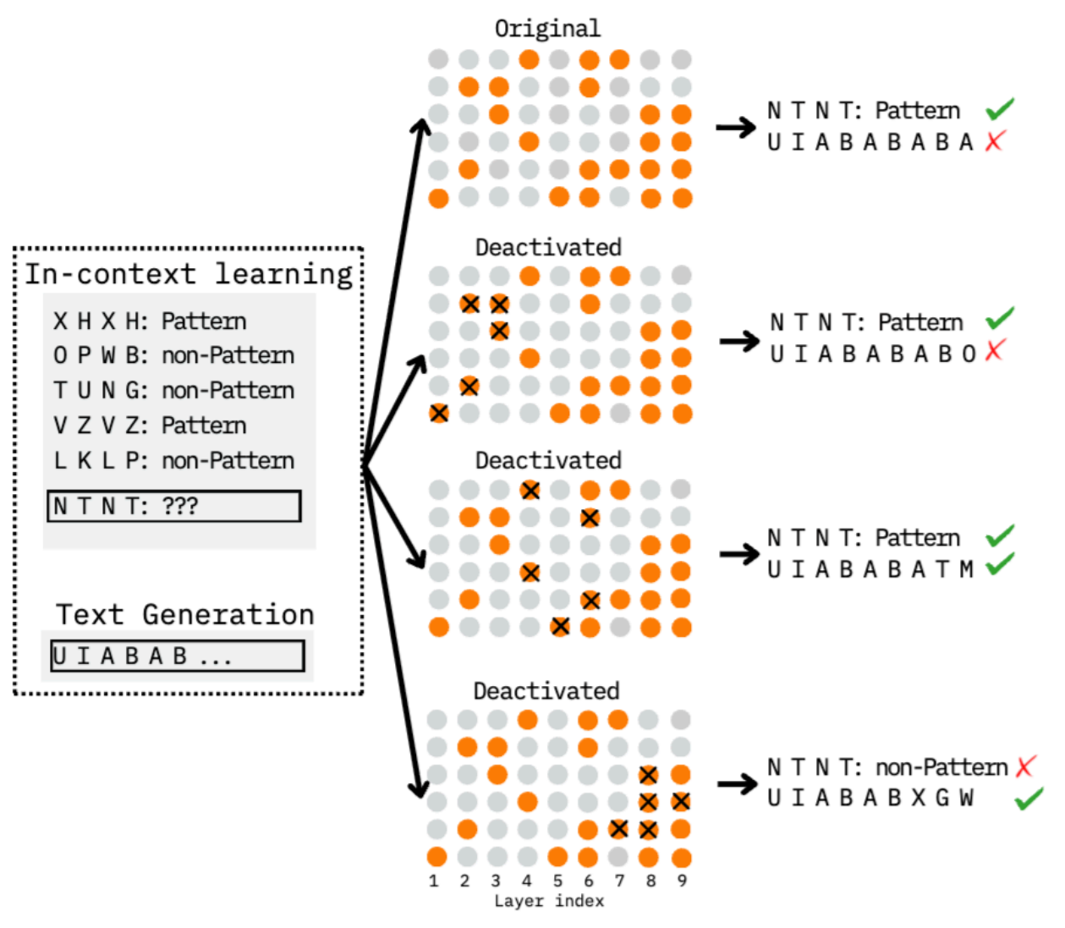
- **Neuronal**: What is the function of individual neurons?
- **Representational**: What information do hidden states encode? How does the model use this information?
- **Algorithmic**: What algorithms and reasoning strategies do models use to generate answers? How are they implemented in which model components?

Neuron-level Analysis

NAACL 2025

Understanding and Controlling Repetition Neurons and Induction Heads in In-context learning

Motivation: In context learning relies on repeated pattern recognition, yet models often generate harmful repetition. The interaction between induction heads and repetition neurons remains unclear.
Contribution: Layer wise ablations reveal complementary roles: induction heads detect patterns, late layer repetition neurons execute them, and removing both breaks in context learning. Targeted ablation of middle layer repetition neurons reduces repetition while largely preserving in context learning.

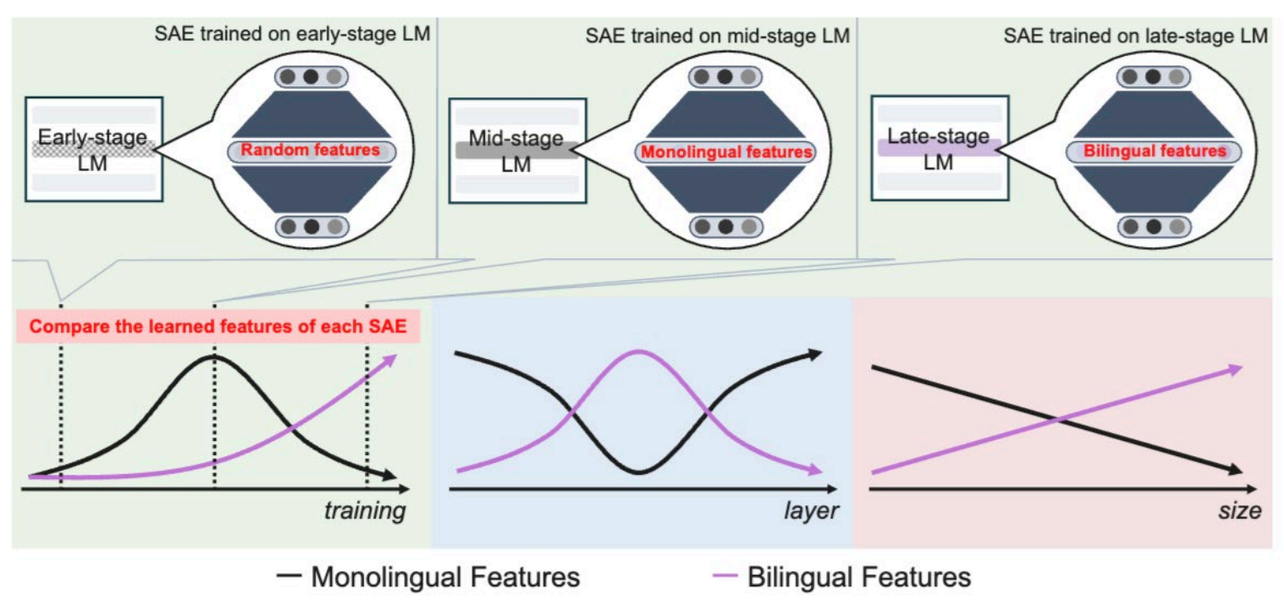


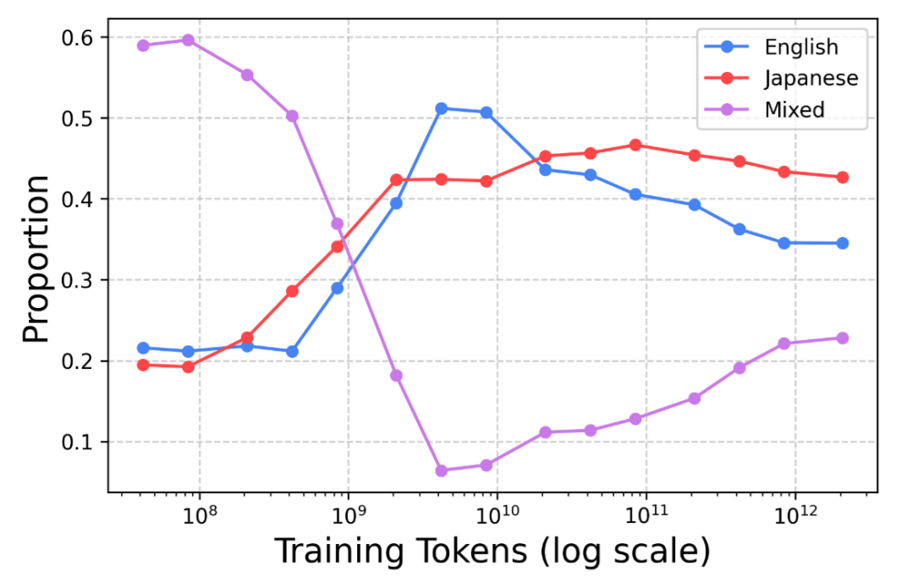
Understanding the Internal Representations of LLMs

EMNLP 2025 Findings

How a Bilingual LM Becomes Bilingual: Tracing Internal Representations with Sparse Autoencoders

Motivation: How bilingual abilities emerge during pretraining is unclear: do models learn languages separately, where alignment forms across layers, and how size affects it.
Contribution: sing sparse autoencoders on English–Japanese LLM-jp checkpoints, we trace feature evolution, show mid-layer alignment, and causally boost mid-training performance by injecting bilingual features.



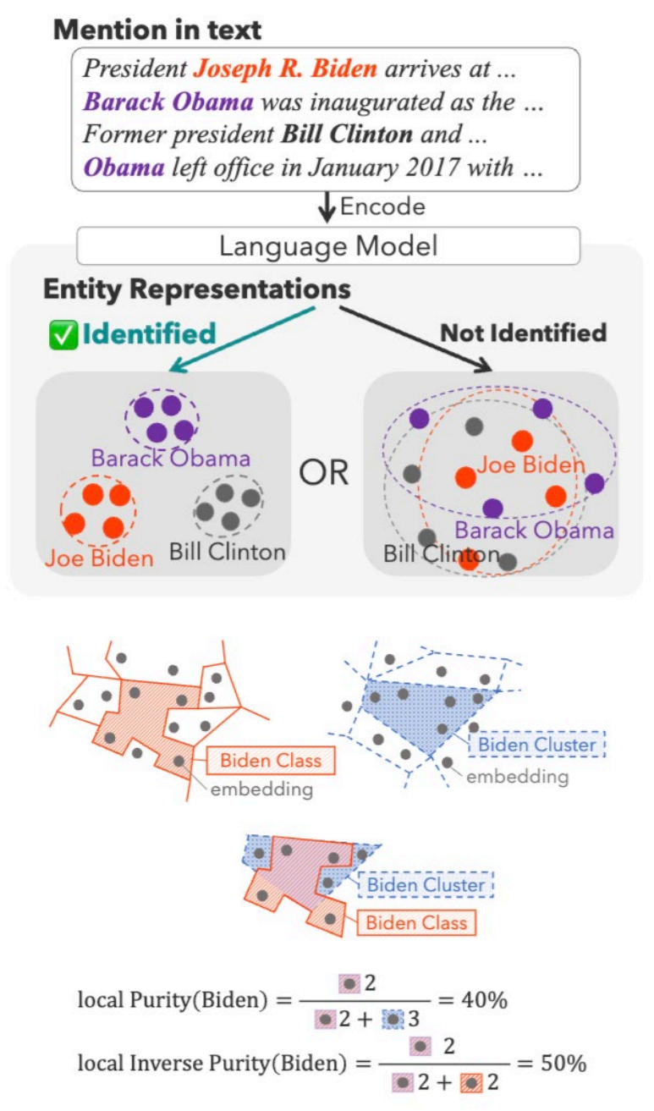


ACL 2025 Findings

Understanding and Controlling Repetition Neurons and Induction Heads in In-context learning

Motivation: Language models store factual knowledge about entities, yet it is unclear whether their internal representations reliably distinguish entities under mention ambiguity and variability.

Contribution: We show that LMs cluster mentions by entity with high precision and recall, encode entity information in low-dimensional early-layer subspaces, and link representation quality to prediction consistency.


Mention in text
President Joseph R. Biden arrives at ...
Barack Obama was inaugurated as the ...
Former president Bill Clinton and ...
Obama left office in January 2017 with ...
Encode
Language Model
Entity Representations
✅Identified
Not Identified
Barack Obama
Joe Biden
Bill Clinton
OR
Joe Biden
Barack Obama
Bill Clinton
Biden Class
embedding
Biden Cluster
embedding
Biden Cluster
Biden Class
local Purity(Biden) = 2 / (2 + 3) = 40%
local Inverse Purity(Biden) = 2 / (2 + 2) = 50%

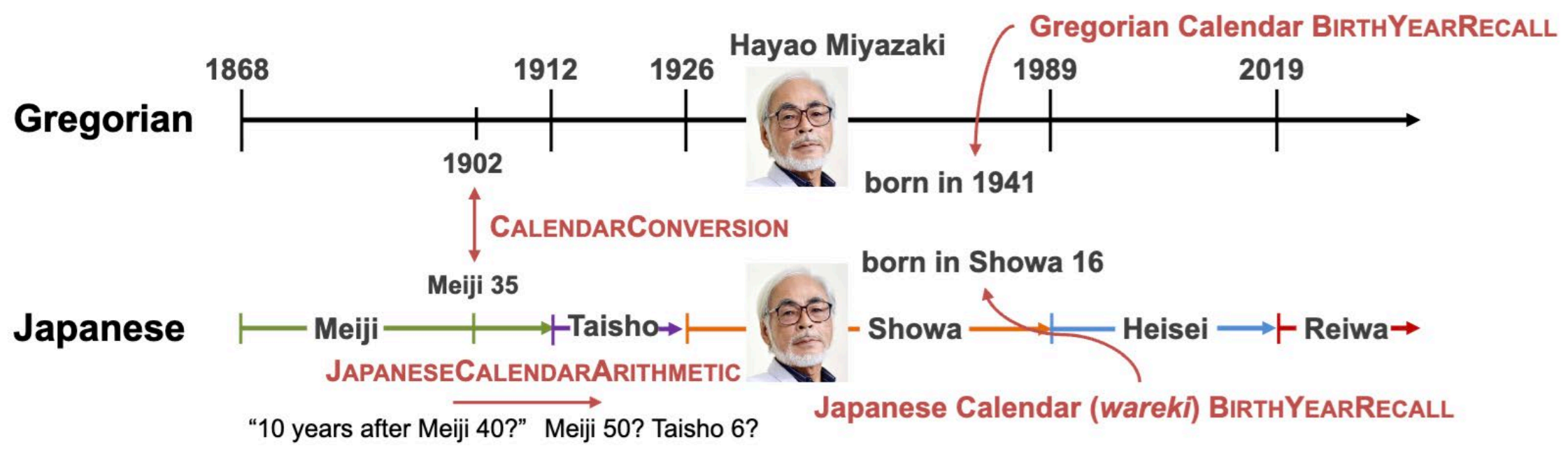
Dissecting the Reasoning Process of LLMs and Vision Models

AACL 2025

Can Language Models Handle a Non-Gregorian Calendar? The Case of the Japanese wareki

Motivation: Language models are evaluated almost exclusively on the Gregorian calendar, so their robustness to non Gregorian systems like Japanese wareki is unclear.

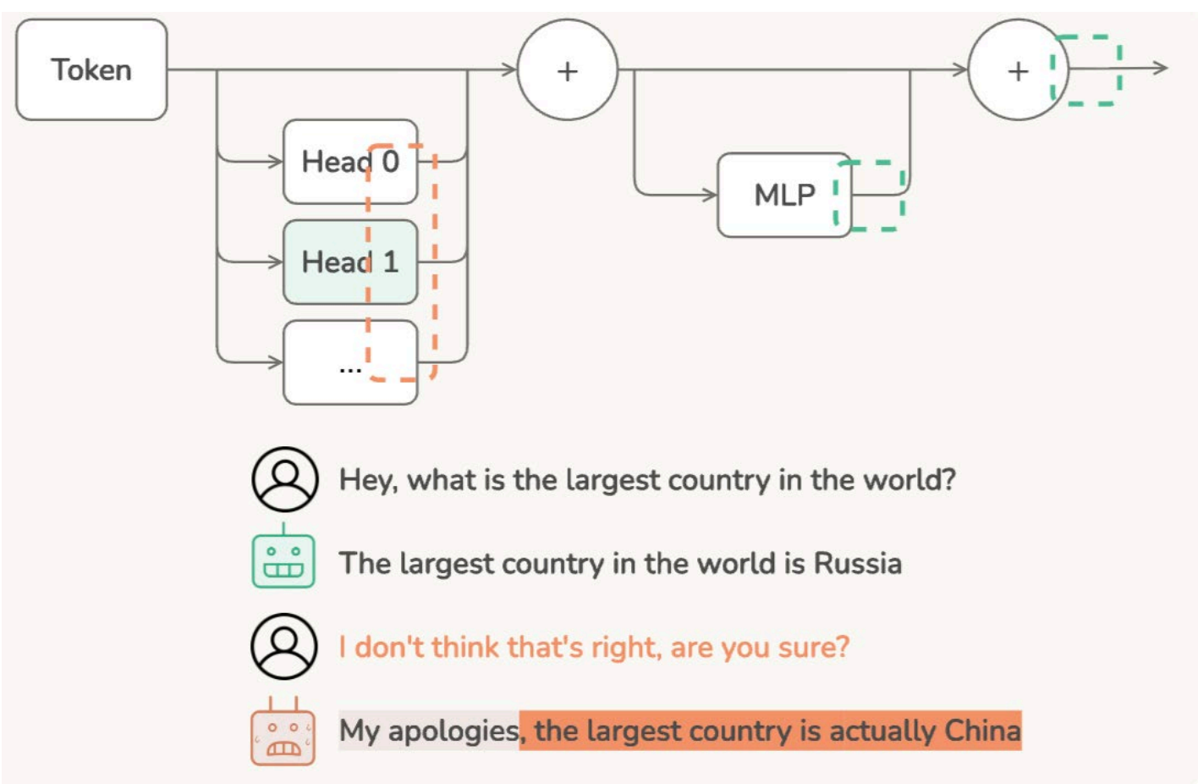
Contribution: We introduce wareki benchmarks for conversion, arithmetic, and birth year recall, showing that even frontier and Japanese centric models fail on arithmetic and recall due to corpus sparsity and Gregorian bias.


Gregorian
1868
1902
1912
1926
Hayao Miyazaki
born in 1941
Gregorian Calendar BirthYearRecall
1989
2019
CalendarConversion
Meiji 35
born in Showa 16
Japanese
Meiji
Taisho
Showa
Heisei
Reiwa
JapaneseCalendarArithmetic
"10 years after Meiji 40?" Meiji 50? Taisho 6?
Japanese Calendar (wareki) BirthYearRecall

EACL 2026

Sycophancy Hides Linearly in the Attention Heads

Motivation: Correct-to-incorrect sycophancy, where a model abandons a correct answer after mild user doubt, harms factual reliability, but it is still an open question which components encode this behavior and where control is most effective.
Contribution: Linear probes localize sycophancy to a sparse set of mid-layer attention heads, where targeted linear steering most effectively reduces reversals, transfers across QA benchmarks, and links the effect to attention on user doubt cues.



Hey, what is the largest country in the world?

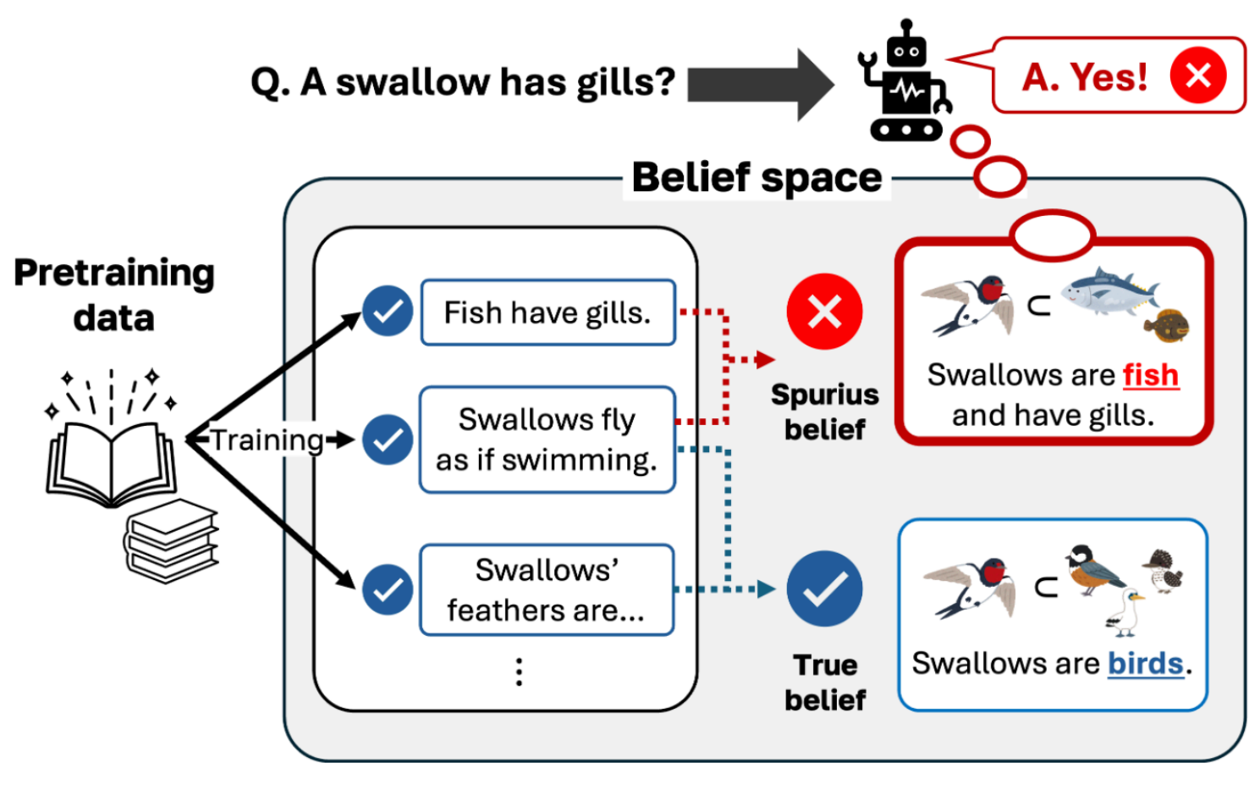
The largest country in the world is Russia

I don't think that's right, are you sure?

My apologies, the largest country is actually China

ACL 2025 Findings

Rectifying Belief Space via Unlearning to Harness LLMs' Reasoning

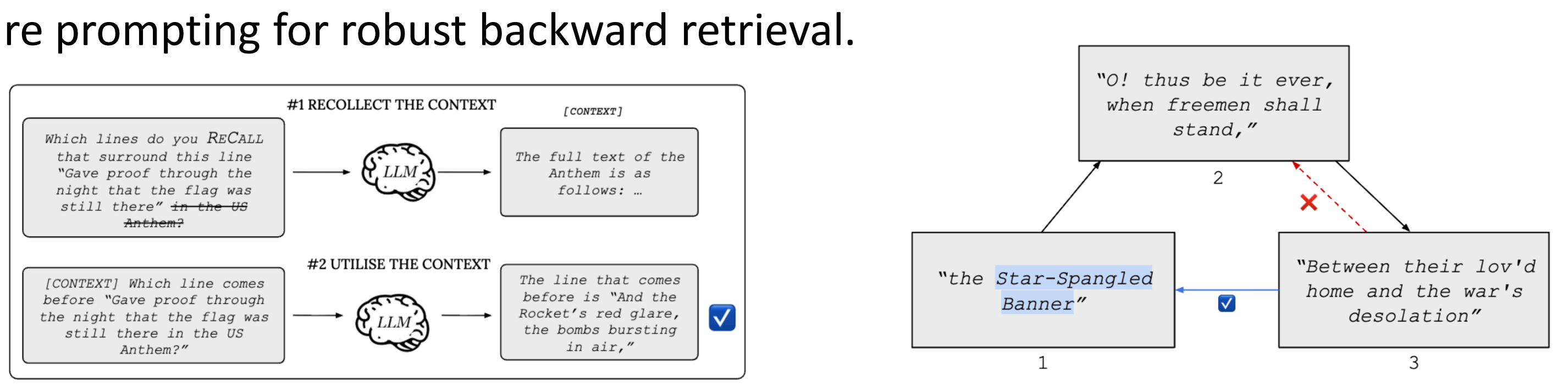


Motivation: LLMs often reason incorrectly despite strong capabilities, possibly due to spurious internal beliefs, but identifying and correcting such beliefs remains poorly understood.
Contribution: We identify beliefs driving answers using forward backward beam search and rectify belief space via unlearning, improving QA accuracy and generalization without harming overall performance.

ACL 2025

Library-Like Behavior In Language Models is Enhanced by Self-Referencing Causal Cycles

Motivation: Autoregressive language models struggle to recall preceding context due to the reversal curse, limiting their effectiveness as reliable memory-like knowledge repositories.
Contribution: We identify self-referencing causal cycles from repeated token sequences, show how they bypass the reversal curse, and demonstrate RECALL-aware prompting for robust backward retrieval.



Which lines do you RECALL that surround this line "Gave proof through the night that the flag was still there" ~~in the US Anthem?~~

The full text of the Anthem is as follows: …

[CONTEXT] Which line comes before "Gave proof through the night that the flag was still there in the US Anthem?"

The line that comes before is "And the Rocket's red glare, the bombs bursting in air,"

"O! thus be it ever, when freemen shall stand,"

"the Star-Spangled Banner"

"Between their lov'd home and the war's desolation"

Publication highlights: Nwadike+ ACL 2025, Niwa+ ACL 2025 Findings, Sakata+ ACL 2025 Findings, Ye+ ACL 2025 Findings, Inaba+ EMNLP 2025 Findings, Ye+ NeurIPS 2025, Nakamura+ 2025 IEEE Access, Sasaki+ AACL 2025, Genadi+ EACL 2026, Kudo+ EACL 2026 Findings, etc.
言語処理学会第31回年次大会優秀賞,スポンサー賞x2