Mitigating Hallucinations in LLMs

Stephen Cowley
Under supervision of Marcus Tomalin
Machine Intelligence Lab

Overview

- Large Language Models (LLMs) hallucination definition

Overview

- Large Language Models (LLMs) hallucination definition
- Investigating combining two existing decoding methods

Overview

- Large Language Models (LLMs) hallucination definition
- Investigating combining two existing decoding methods
- Investigating a modification to existing method

Overview

- Large Language Models (LLMs) hallucination definition
- Investigating combining two existing decoding methods
- Investigating a modification to existing method
- New method to alter LLM “thought process”

LLM Hallucination

- Definitions vary

LLM Hallucination

- Definitions vary
- When “a model makes factual errors” – OpenAI

LLM Hallucination

- Definitions vary
- When “a model makes factual errors” – OpenAI
- Hallucination rate ≠ 1 - accuracy

LLM Hallucination

- Definitions vary
- When “a model makes factual errors” – OpenAI
- Hallucination rate ≠ 1 - accuracy
- But impractical to label

LLM Hallucination

- Definitions vary
- When “a model makes factual errors” – OpenAI
- Hallucination rate ≠ 1 - accuracy
- But impractical to label
- Objective: improve factual reliability and reasoning

Context-Aware Decoding (CAD)

- Reminder – Large Language Model (LLM)

$$\mathrm{P}(w_t \mid \mathbf{w}_{<t})$$

$$\mathbf{w}_{<t} = [w_1, w_2, \ldots, w_{t-1}]$$

Next token probabilities,
given history

Context-Aware Decoding (CAD)

- Reminder – Large Language Model (LLM)

$$\mathrm{P}(w_t \mid \mathbf{w}_{<t}) \qquad \mathbf{w}_{<t} = [w_1, w_2, \ldots, w_{t-1}]$$

Next token probabilities, given history

- **CAD is:**

(prompt $\mathbf{x}$, sequence so far $\mathbf{y}_{<t}$)

Context-Aware Decoding (CAD)

- Reminder – Large Language Model (LLM)

$$\mathrm{P}(w_t \mid \mathbf{w}_{<t}) \qquad \mathbf{w}_{<t} = [w_1, w_2, \ldots, w_{t-1}]$$

Next token probabilities,
given history

- **CAD is:**

(prompt $\mathbf{x}$, sequence so far $\mathbf{y}_{<t}$)

$$\mathrm{P}_{\mathrm{CAD}}(y_t) \propto \mathrm{P}(y_t \mid \mathbf{c}, \mathbf{x}, \mathbf{y}_{<t}) \left(\frac{\mathrm{P}(y_t \mid \mathbf{c}, \mathbf{x}, \mathbf{y}_{<t})}{\mathrm{P}(y_t \mid \mathbf{x}, \mathbf{y}_{<t})} \right)^{\alpha}$$

Context-Aware Decoding (CAD)

- Reminder – Large Language Model (LLM)

$$\mathrm{P}(w_t \mid \mathbf{w}_{<t}) \qquad \mathbf{w}_{<t} = [w_1, w_2, \ldots, w_{t-1}]$$

Next token probabilities,
given history

- **CAD is:**

(prompt $\mathbf{x}$, sequence so far $\mathbf{y}_{<t}$)

$$\mathrm{P}_{\mathrm{CAD}}(y_t) \propto \mathrm{P}(y_t \mid \underset{\textbf{Context}}{\boxed{\mathbf{c},}} \mathbf{x}, \mathbf{y}_{<t}) \left(\frac{\mathrm{P}(y_t \mid \overset{\textbf{Context}}{\boxed{\mathbf{c},}} \mathbf{x}, \mathbf{y}_{<t})}{\mathrm{P}(y_t \mid \mathbf{x}, \mathbf{y}_{<t})} \right)^{\alpha}$$

Decoding by Contrasting Layers (DoLa)

- Also contrasts distributions

Decoding by Contrasting Layers (DoLa)

- Also contrasts distributions
- Transformer layers are internal model states

Decoding by Contrasting Layers (DoLa)

- Also contrasts distributions
- Transformer layers are internal model states
- **DoLa is:**

$$\mathrm{P}_{\mathrm{DoLa}}(y_t) \propto \mathrm{P}(y_t)\left(\frac{\mathrm{P}(y_t)}{\boxed{\mathrm{P}^{(l)}(y_t)}}\right)$$

Distribution at previous layer

Results

1. MemoTrap

Context:

Write a quote that ends in the word "early"

Prompt:

Better late than

Results

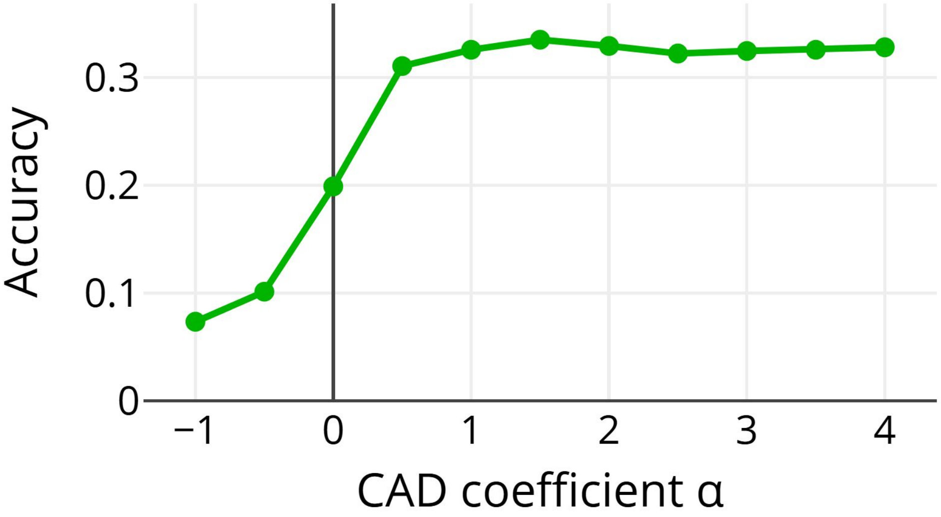
1. MemoTrap

Context:

Write a quote that ends in the word "early"

Prompt:

Better late than



Results

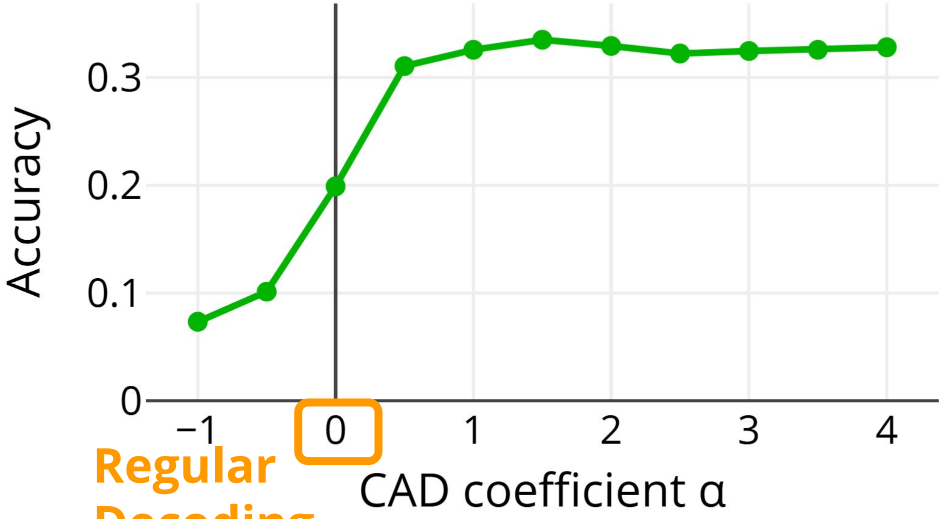
1. MemoTrap

Context:

Write a quote that ends in the word “early”

Prompt:

Better late than

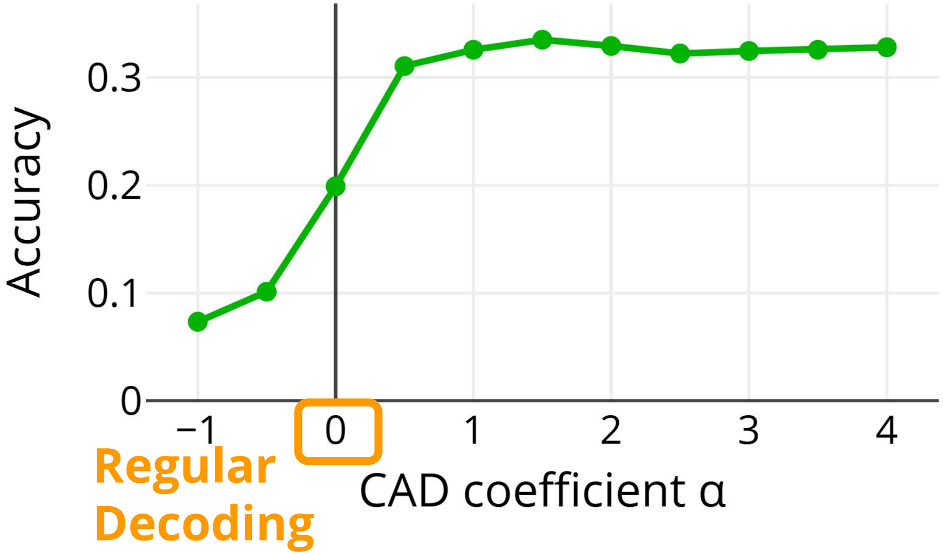


Regular Decoding

Results

1. MemoTrap

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



Regular Decoding

2. Natural Questions

Context:
Ashrita Furman (born Keith Furman, September 16, 1954) is a Guinness World Records record-breaker. …
Prompt:
who holds the world record for the most world records?

Results

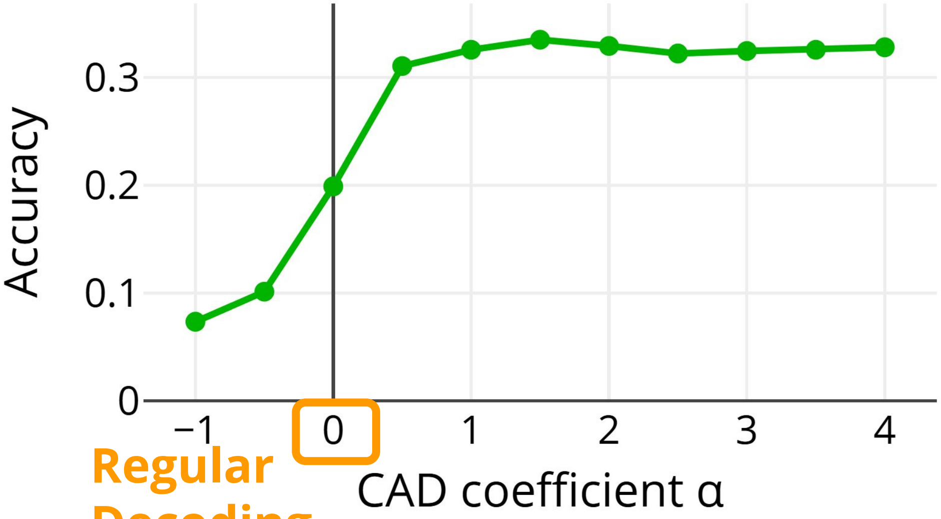
1. MemoTrap

Context:

Write a quote that ends in the word "early"

Prompt:

Better late than



Regular Decoding

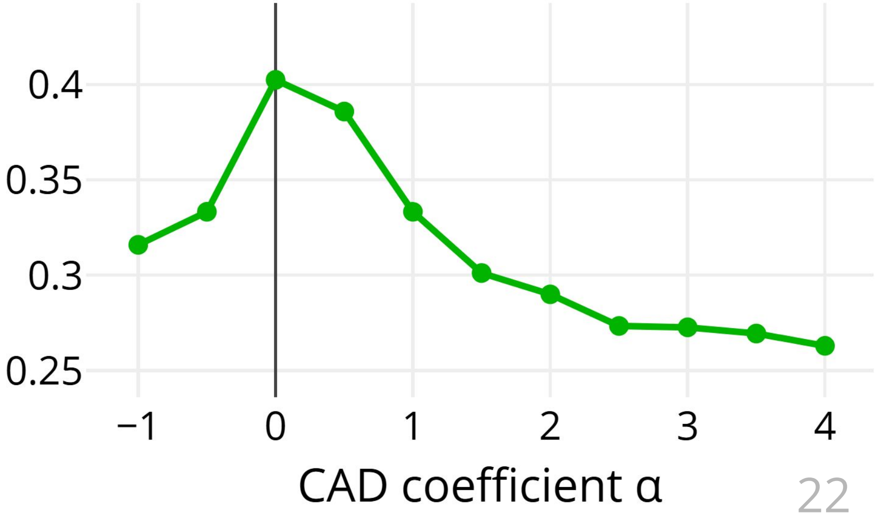
2. Natural Questions

Context:

Ashrita Furman (born Keith Furman, September 16, 1954) is a Guinness World Records record-breaker. …

Prompt:

who holds the world record for the most world records?

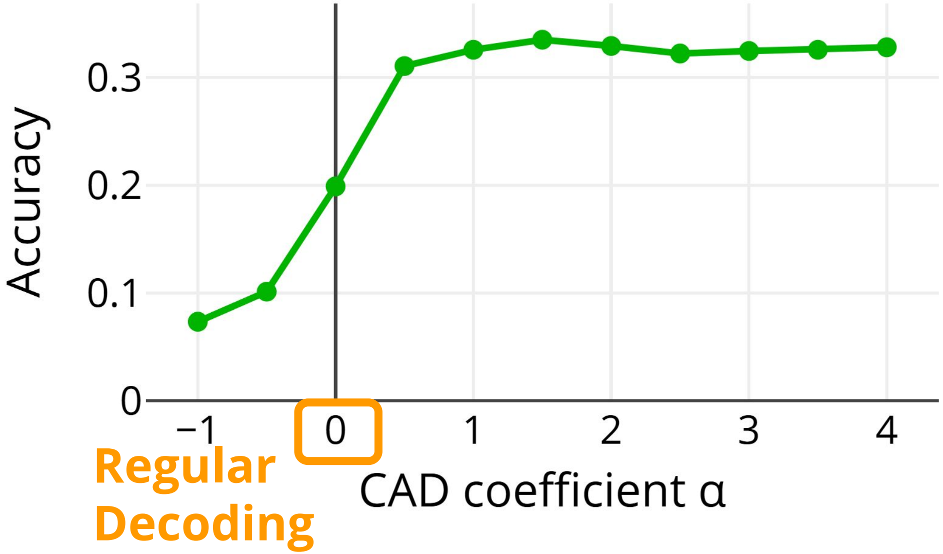


Results

1. MemoTrap

Context:
Write a quote that ends in the word "early"

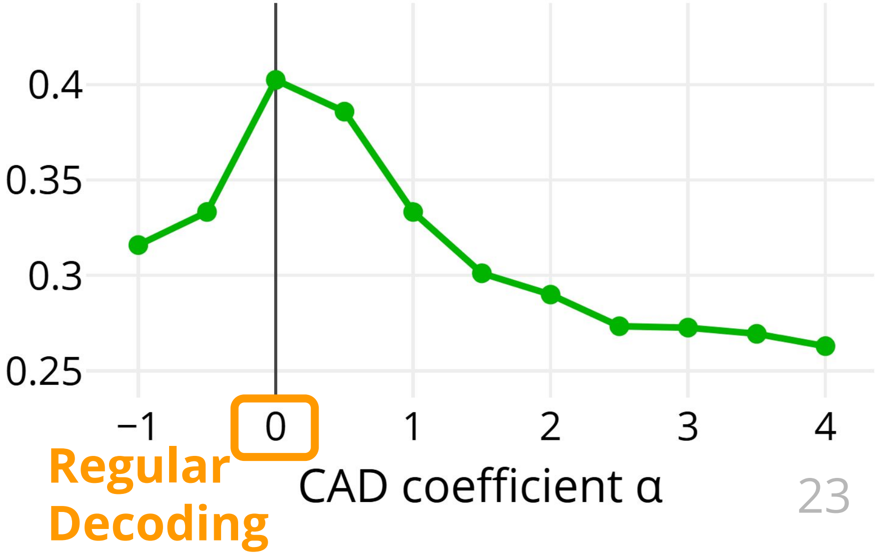
Prompt:
Better late than



2. Natural Questions

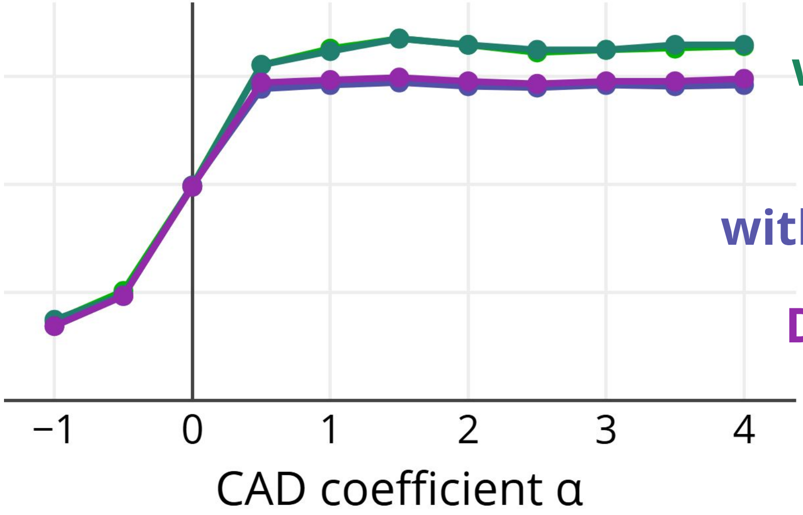
Context:
Ashrita Furman (born Keith Furman, September 16, 1954) is a Guinness World Records record-breaker. …

Prompt:
who holds the world record for the most world records?



Results

1. MemoTrap



Normal

DoLa on with-context

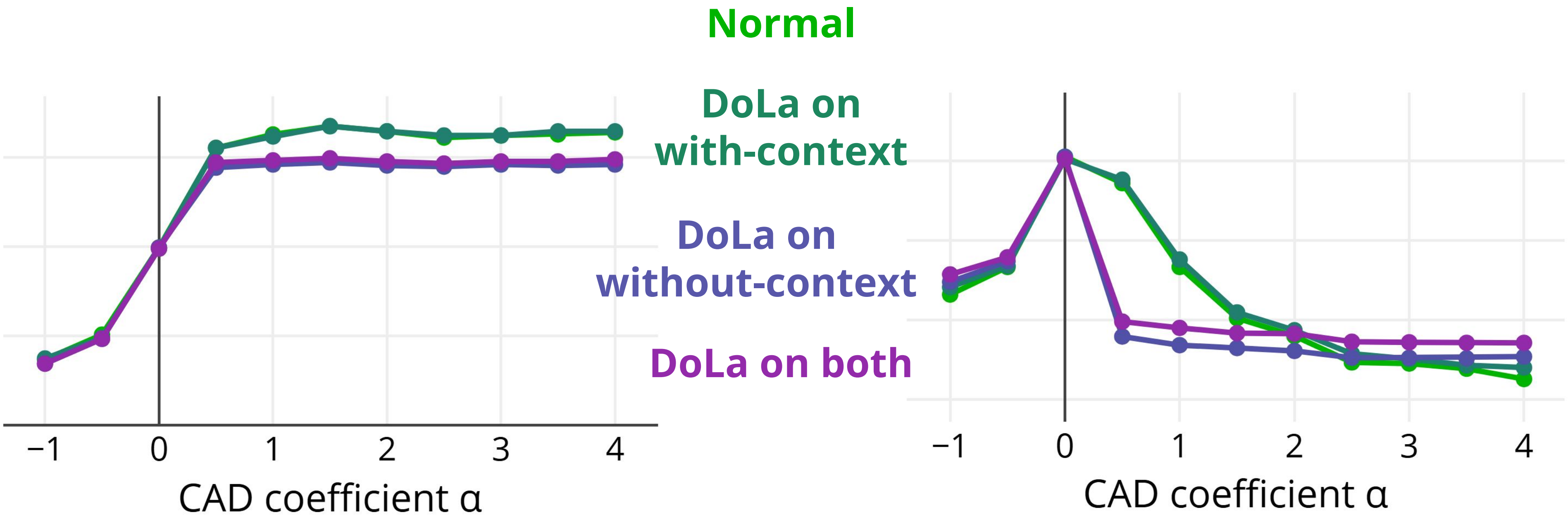
DoLa on without-context

DoLa on both

Results

1. MemoTrap

2. Natural Questions

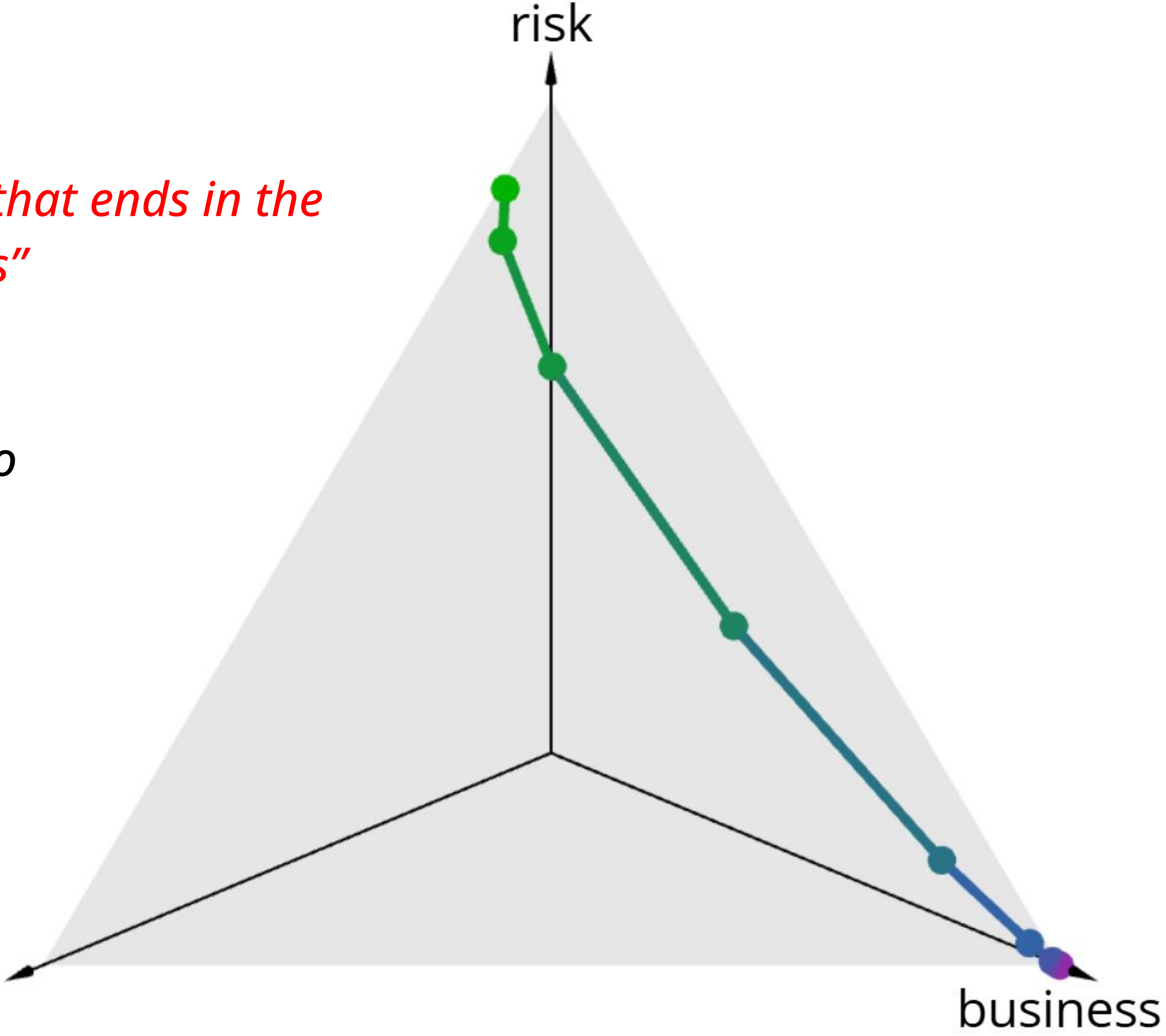


Context:

Write a quote that ends in the word "business"

Prompt:

Advisers run no

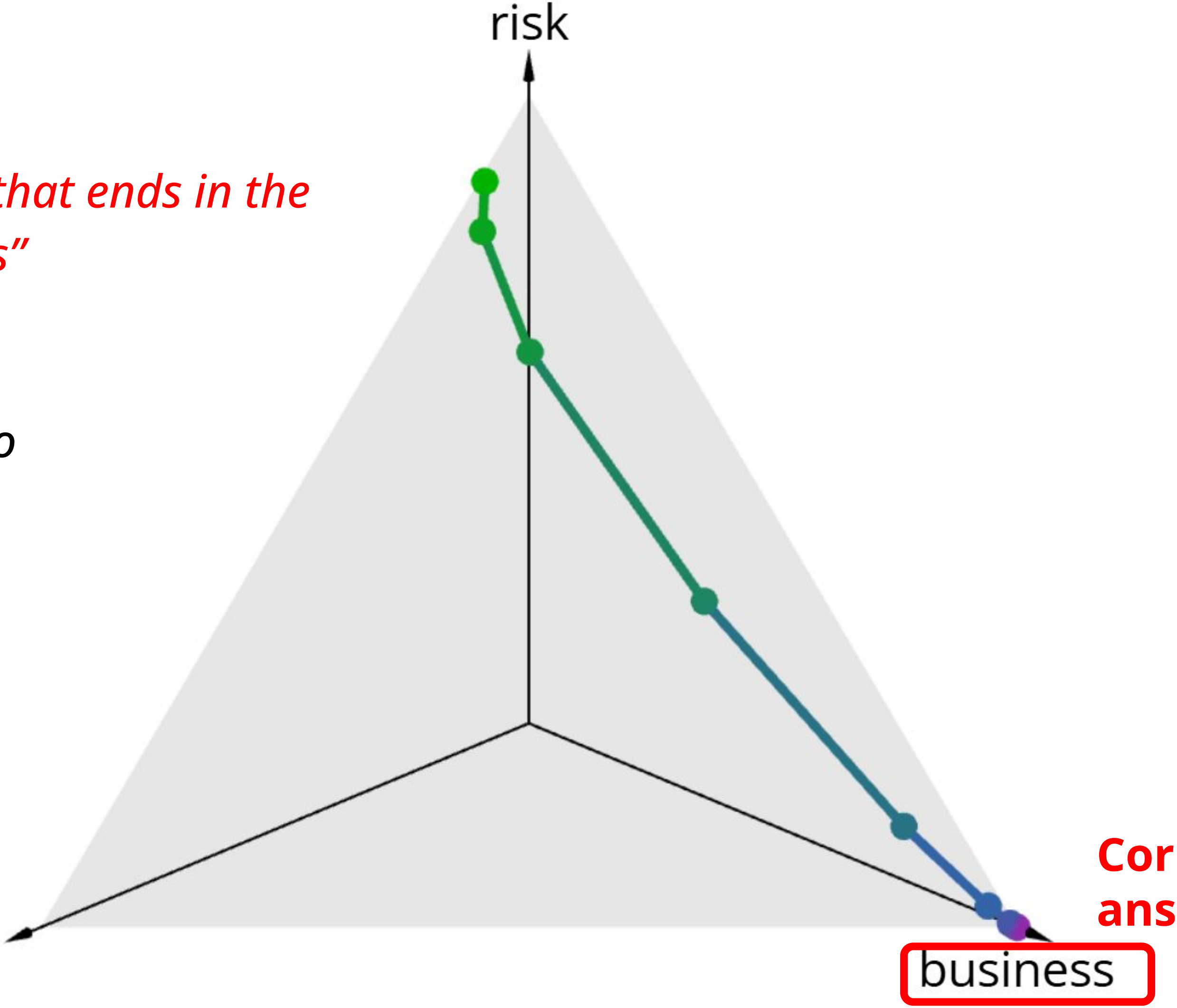


Context:

Write a quote that ends in the word "business"

Prompt:

Advisers run no



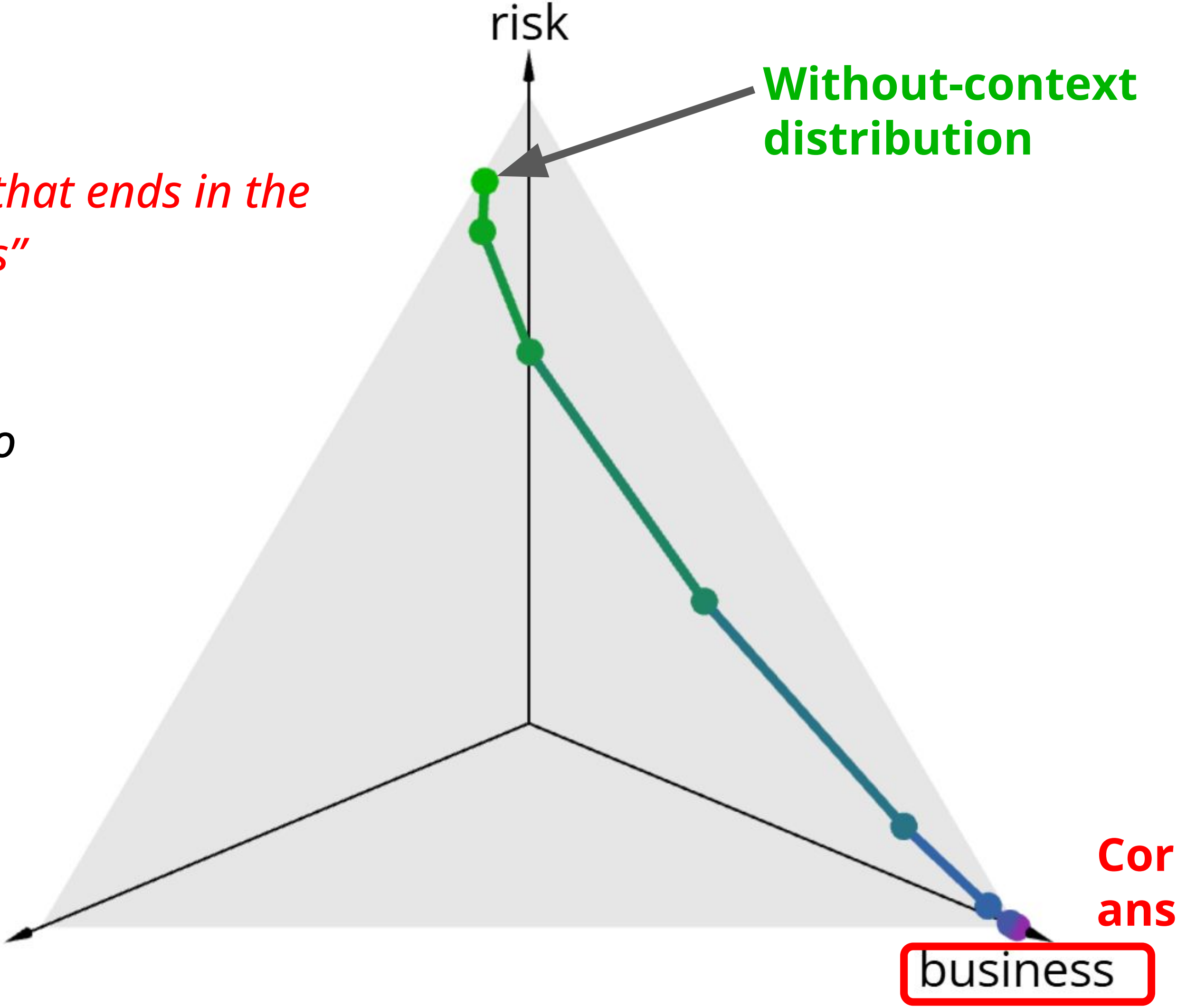
Correct answer

Context:

Write a quote that ends in the word "business"

Prompt:

Advisers run no



Without-context distribution

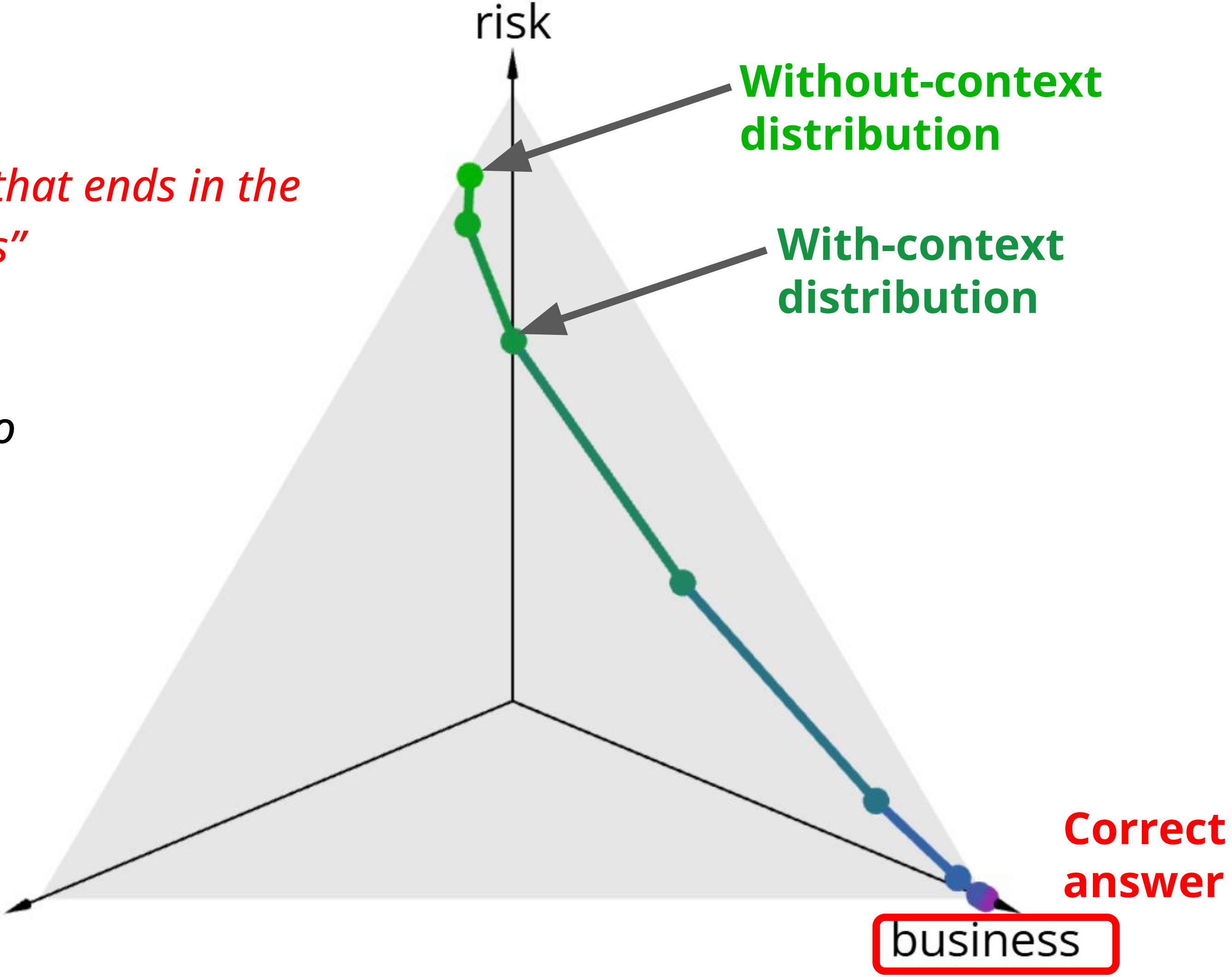
Correct answer

Context:

Write a quote that ends in the word "business"

Prompt:

Advisers run no

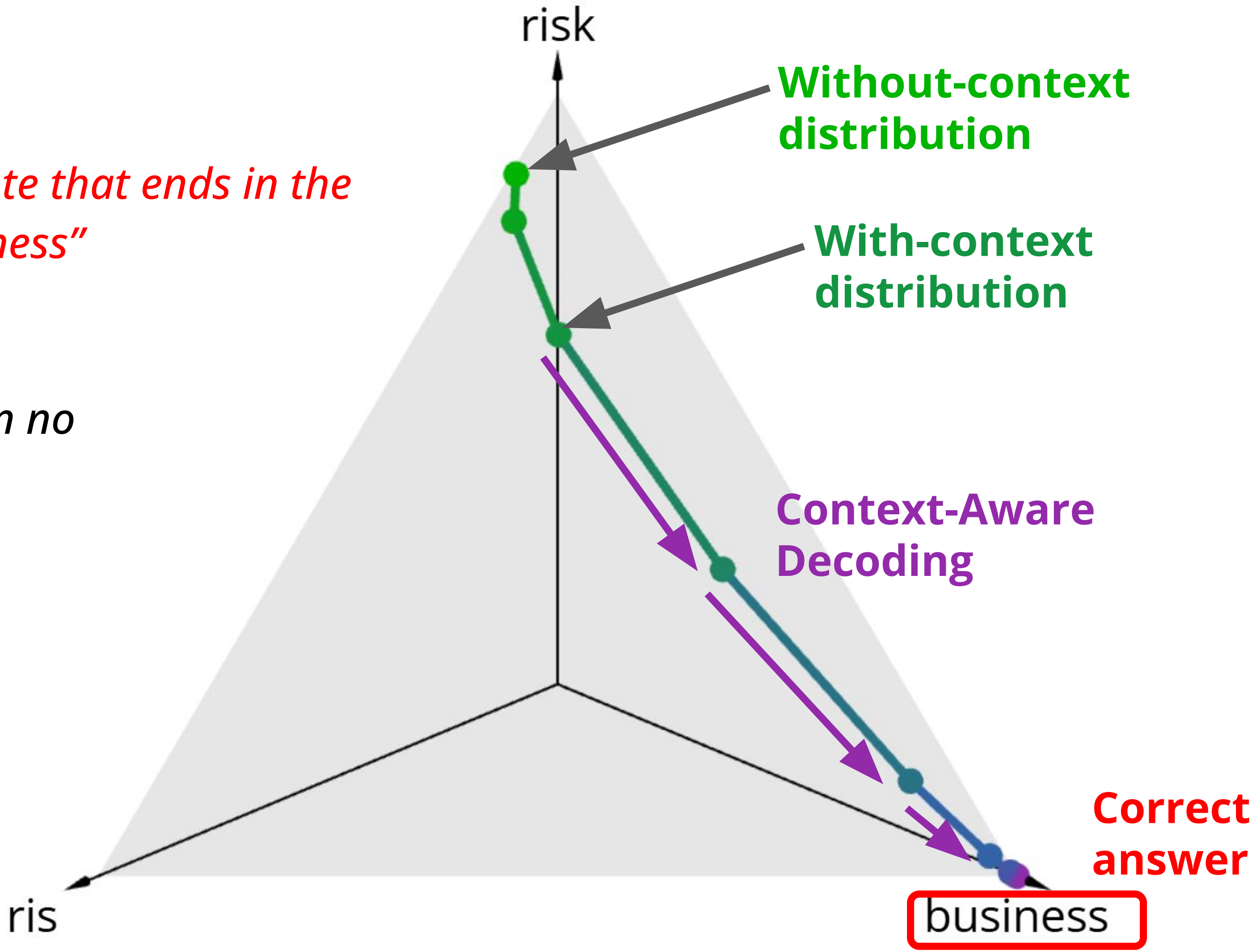


Context:

Write a quote that ends in the word “business”

Prompt:

Advisers run no

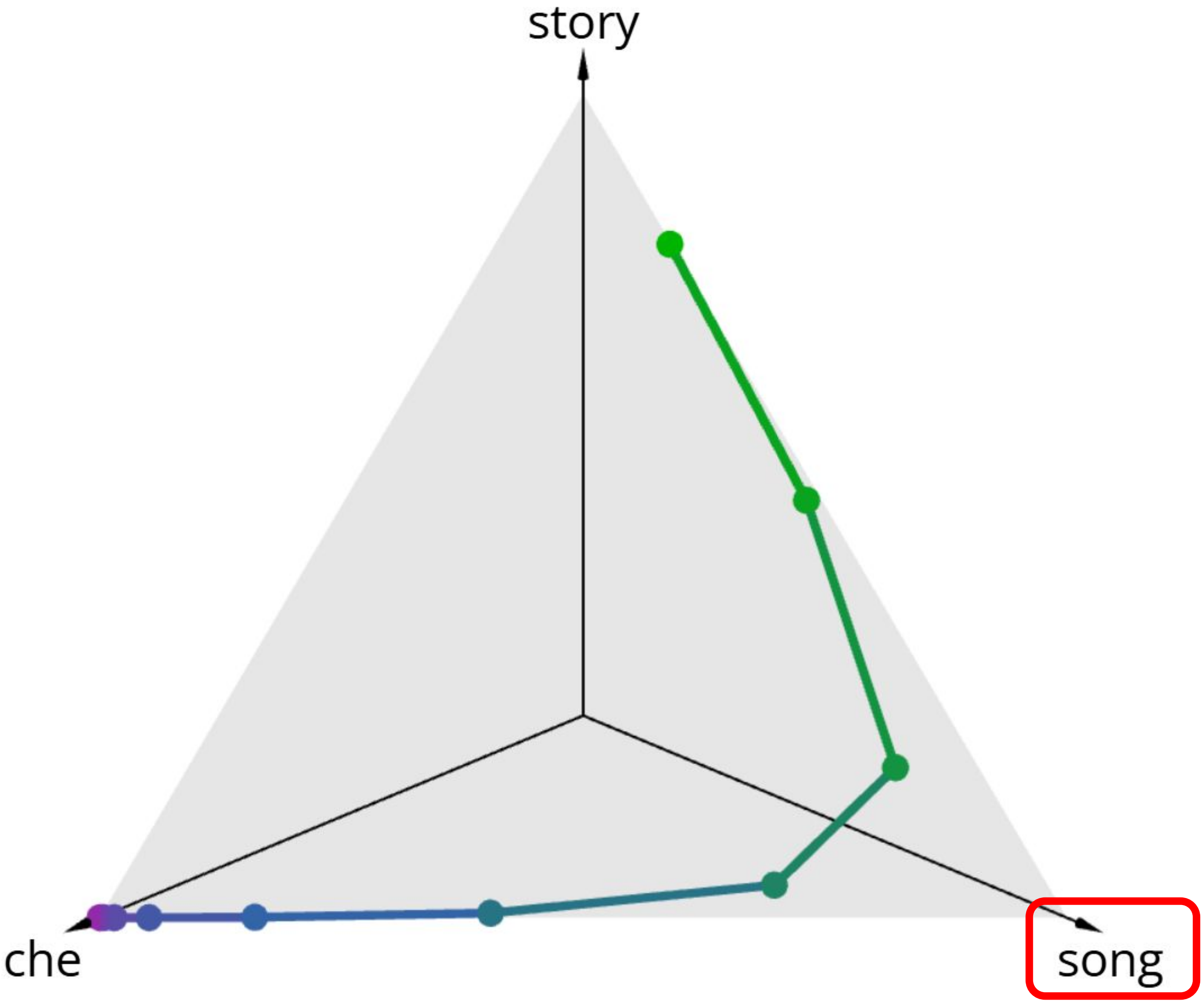


Without-context distribution

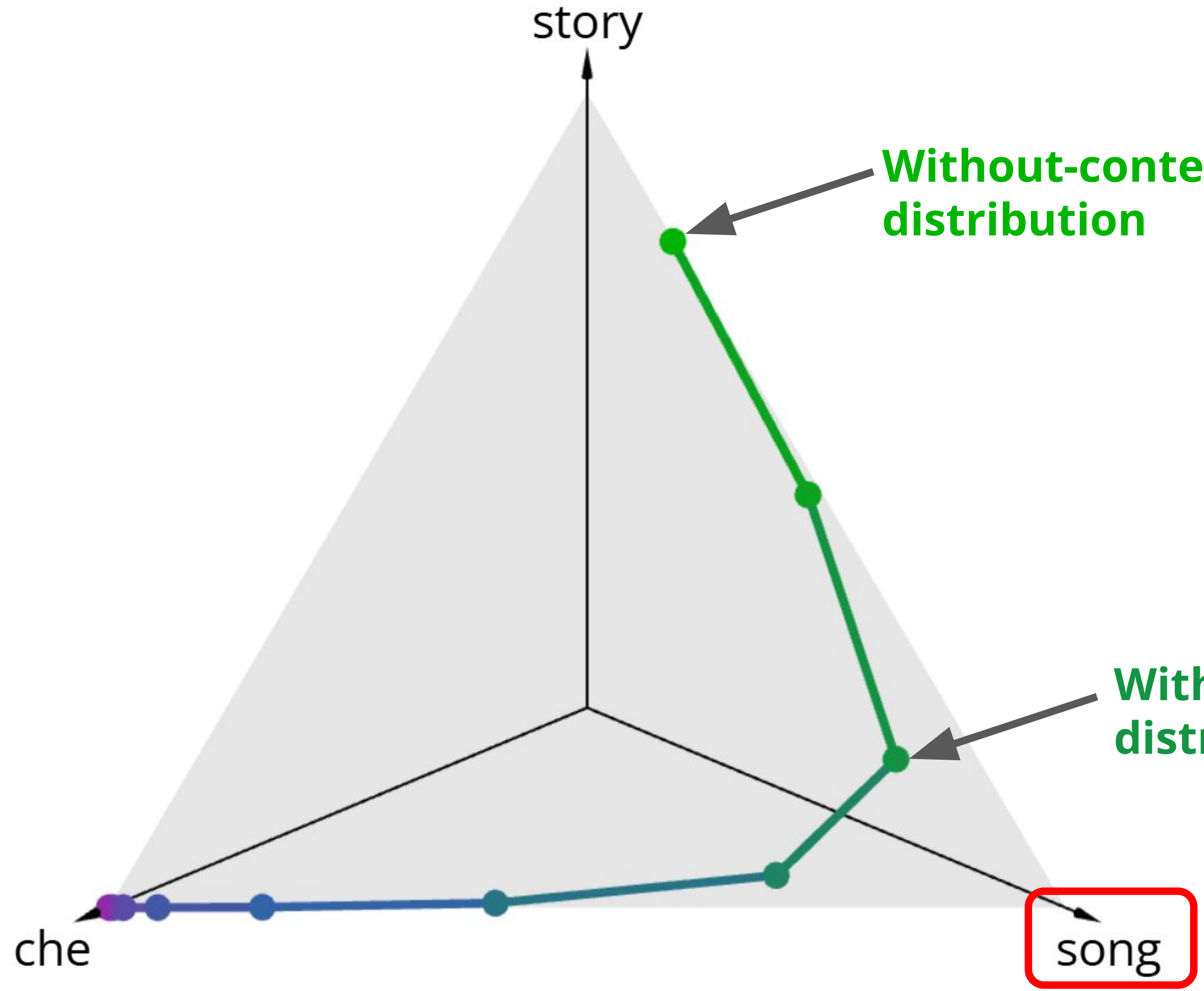
With-context distribution

Context-Aware Decoding

Correct answer



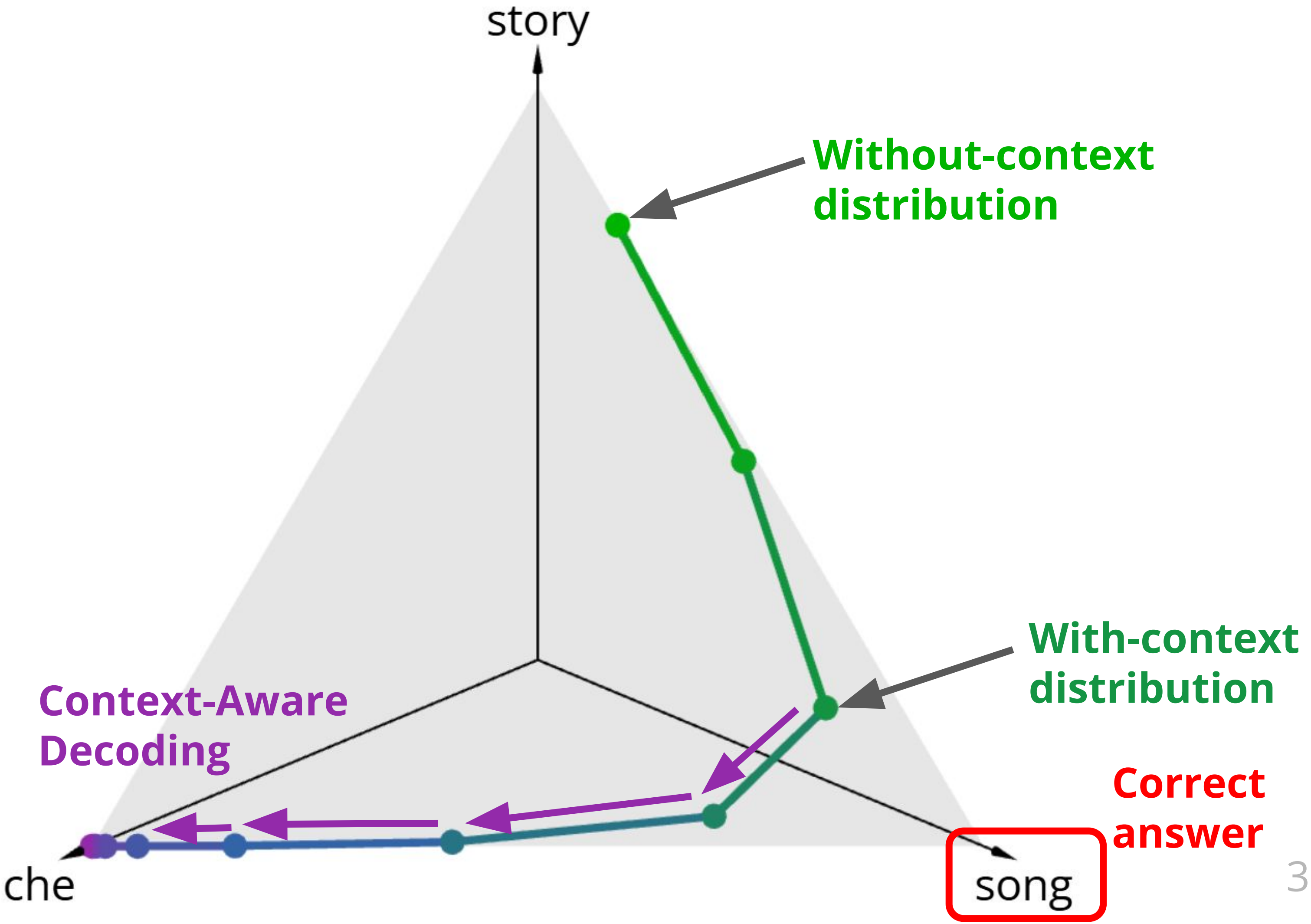
Correct answer



Without-context distribution

With-context distribution

Correct answer

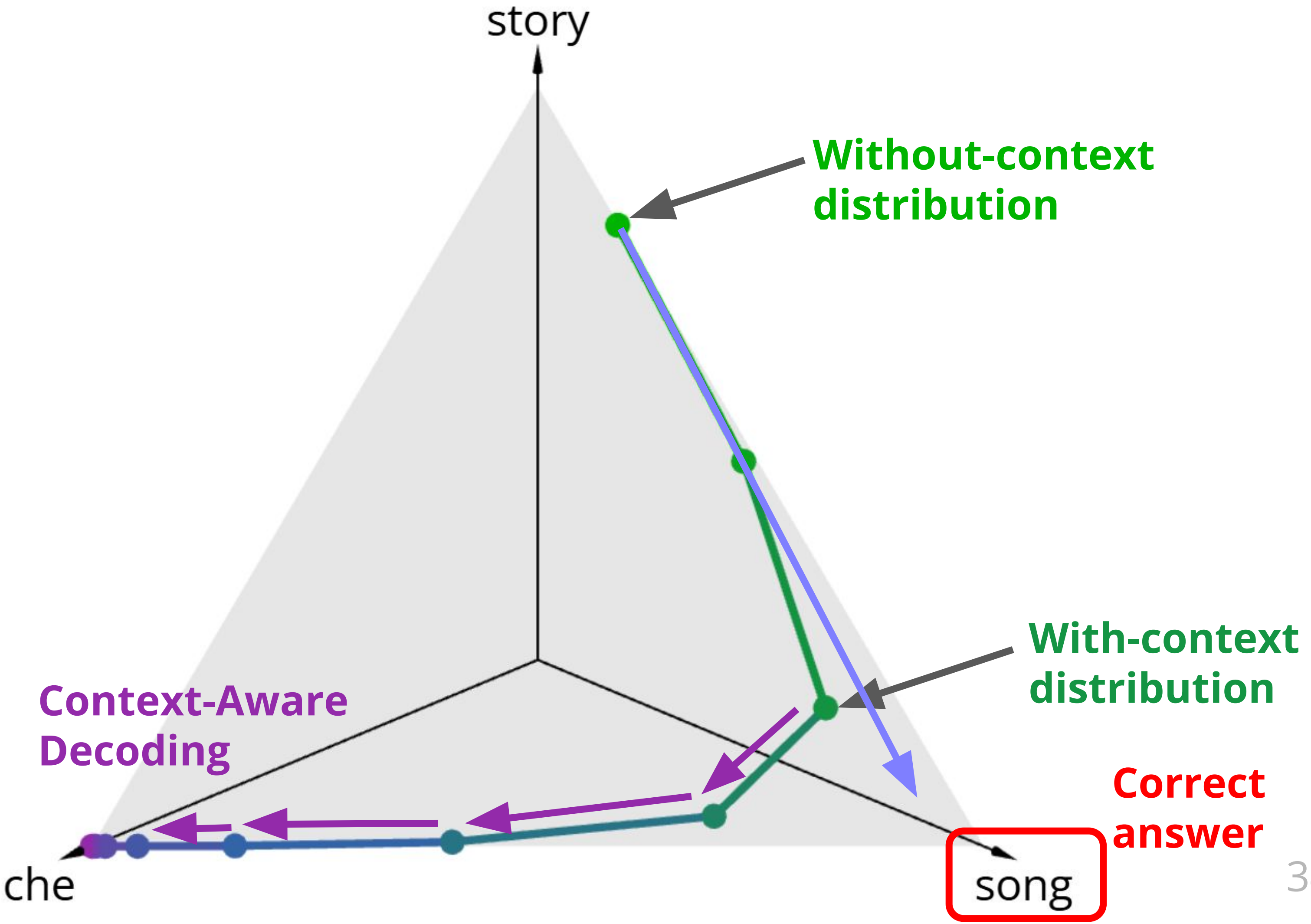


Without-context distribution

With-context distribution

Context-Aware Decoding

Correct answer


story
Without-context distribution
With-context distribution
Context-Aware Decoding
Correct answer
che
song

Additive Context-Aware Decoding

Add the difference instead of multiplying by ratio:

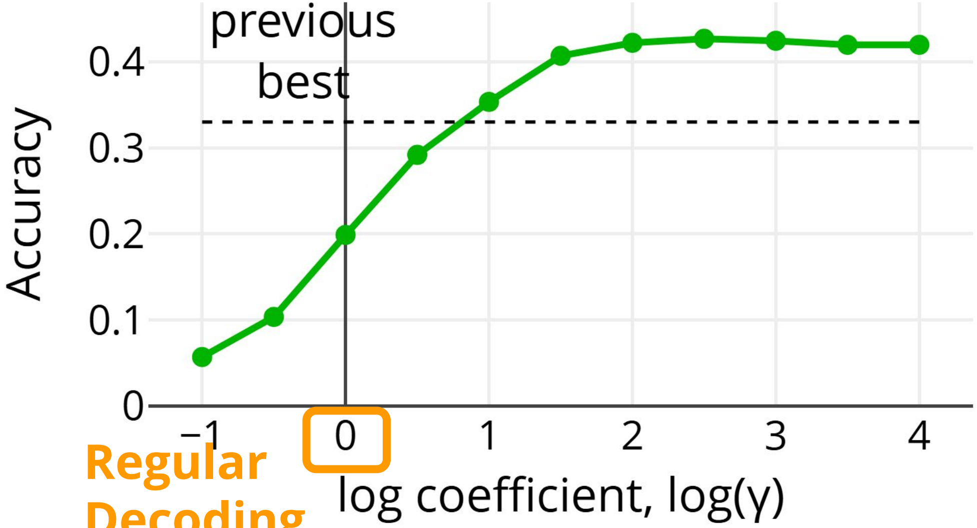
$$\mathrm{P}'_{\mathrm{AddCAD}}(y_t) = \mathrm{P}(y_t \mid \mathbf{x}, \mathbf{y}_{<t}) \ + \gamma\left(\mathrm{P}(y_t \mid \mathbf{c}, \mathbf{x}, \mathbf{y}_{<t}) - \mathrm{P}(y_t \mid \mathbf{x}, \mathbf{y}_{<t})\right)$$

Additive Context-Aware Decoding

Add the difference instead of multiplying by ratio:

$$\mathrm{P}'_{\mathrm{AddCAD}}(y_t) = \mathrm{P}(y_t \mid \mathbf{x}, \mathbf{y}_{<t}) \; + \gamma\left(\mathrm{P}(y_t \mid \mathbf{c}, \mathbf{x}, \mathbf{y}_{<t}) - \mathrm{P}(y_t \mid \mathbf{x}, \mathbf{y}_{<t})\right)$$

1. MemoTrap – even better



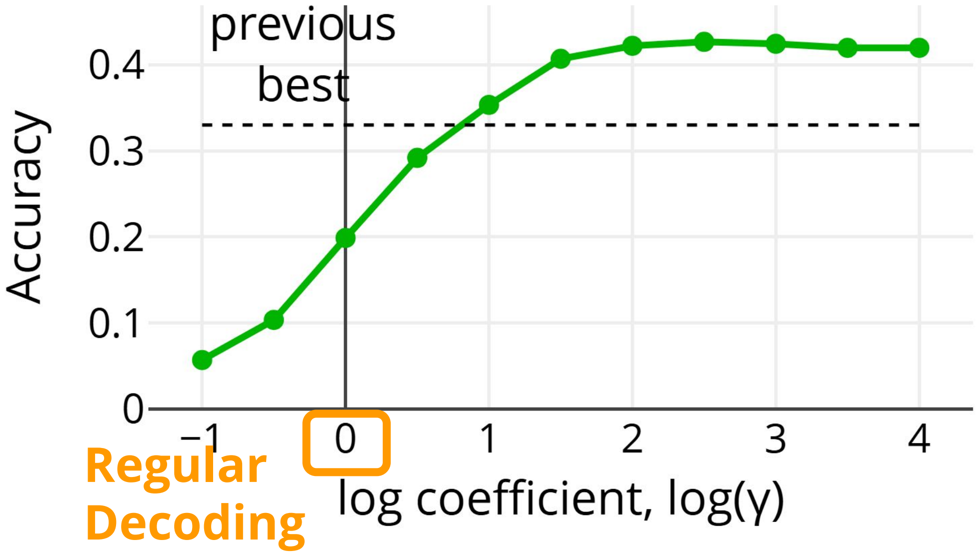
Regular Decoding

Additive Context-Aware Decoding

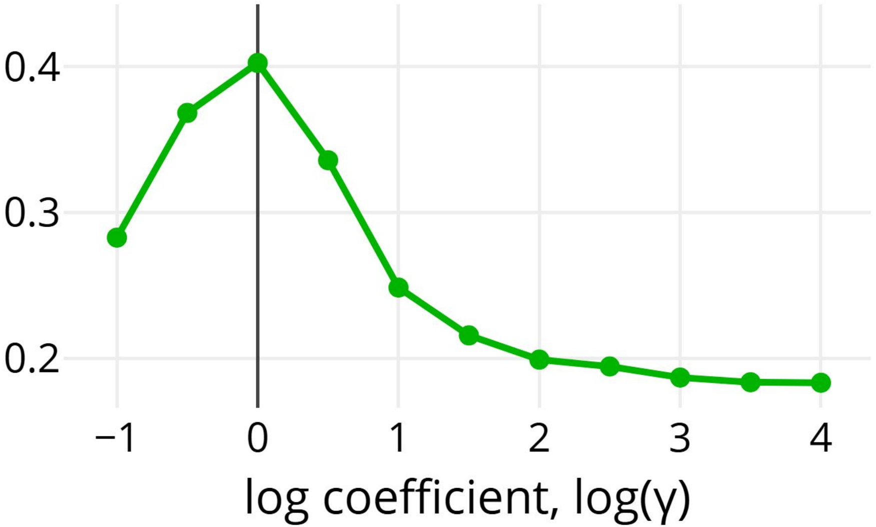
Add the difference instead of multiplying by ratio:

$$\mathrm{P}'_{\mathrm{AddCAD}}(y_t) = \mathrm{P}(y_t \mid \mathbf{x}, \mathbf{y}_{<t}) \; + \gamma\,(\mathrm{P}(y_t \mid \mathbf{c}, \mathbf{x}, \mathbf{y}_{<t}) - \mathrm{P}(y_t \mid \mathbf{x}, \mathbf{y}_{<t}))$$

1. MemoTrap – even better



2. Natural Questions – even worse



Doubt Injection

- Chain-of-Thought is a reasoning process

Doubt Injection

- Chain-of-Thought is a reasoning process
- Chain-of-Thought works

Doubt Injection

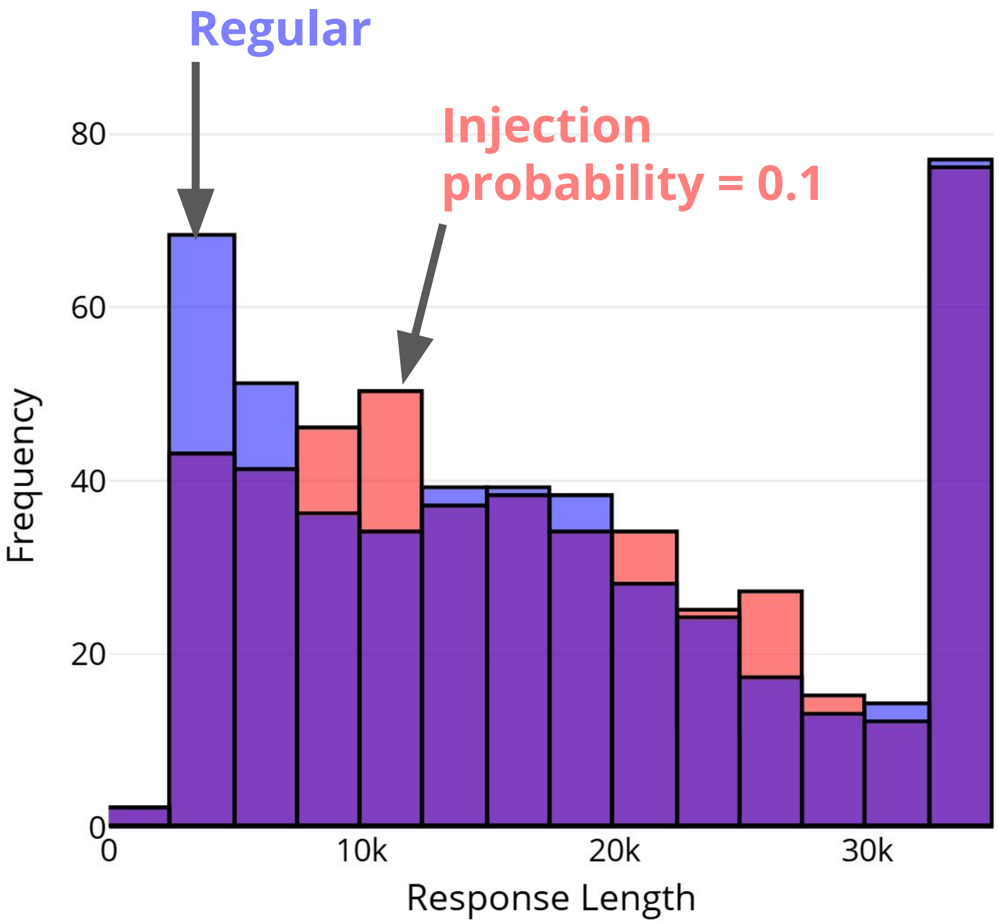
- Chain-of-Thought is a reasoning process
- Chain-of-Thought works
- Explore more ideas

Doubt Injection

- Chain-of-Thought is a reasoning process
- Chain-of-Thought works
- Explore more ideas
- Try randomly inject e.g. “But” at new paragraph

Doubt Injection

- Chain-of-Thought is a reasoning process
- Chain-of-Thought works
- Explore more ideas
- Try randomly inject e.g. “But” at new paragraph
- Tends to increase response length

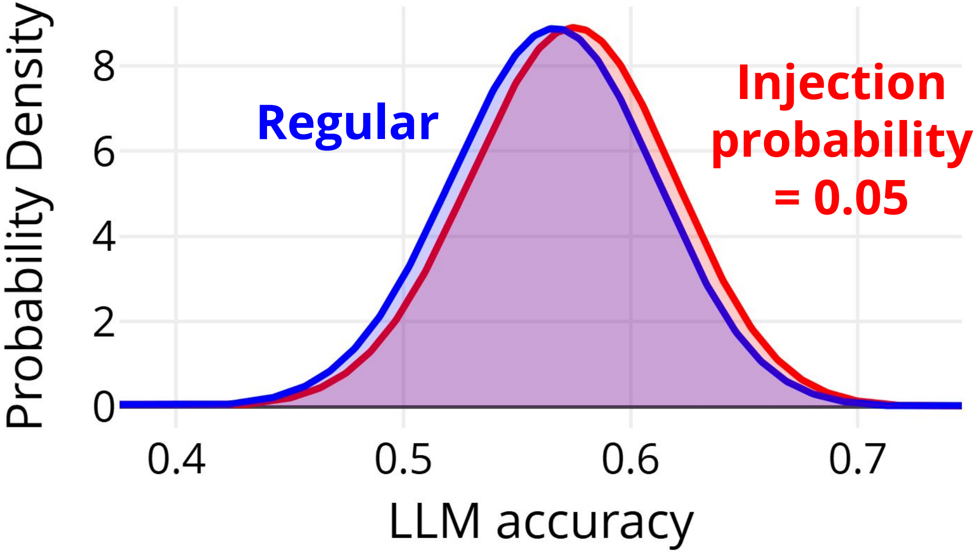


Doubt Injection – Results

- Adversarial questions: accuracy rise 26.1% → 26.7%
- Arithmetic reasoning: accuracy rise 56.7% → 57.5%

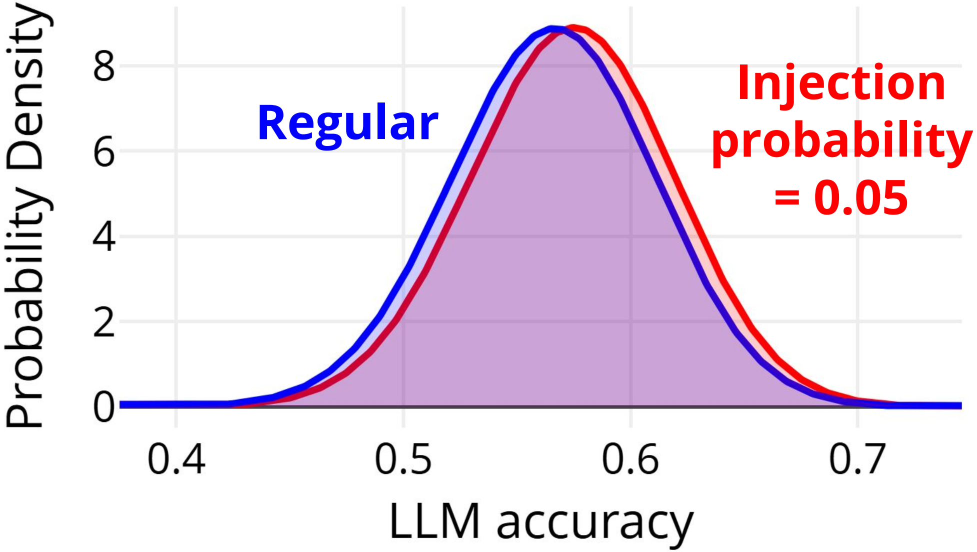
Doubt Injection – Results

- Adversarial questions: accuracy rise 26.1% → 26.7%
- Arithmetic reasoning: accuracy rise 56.7% → 57.5%
- Statistically insignificant (~58%, 55% probability it helps)



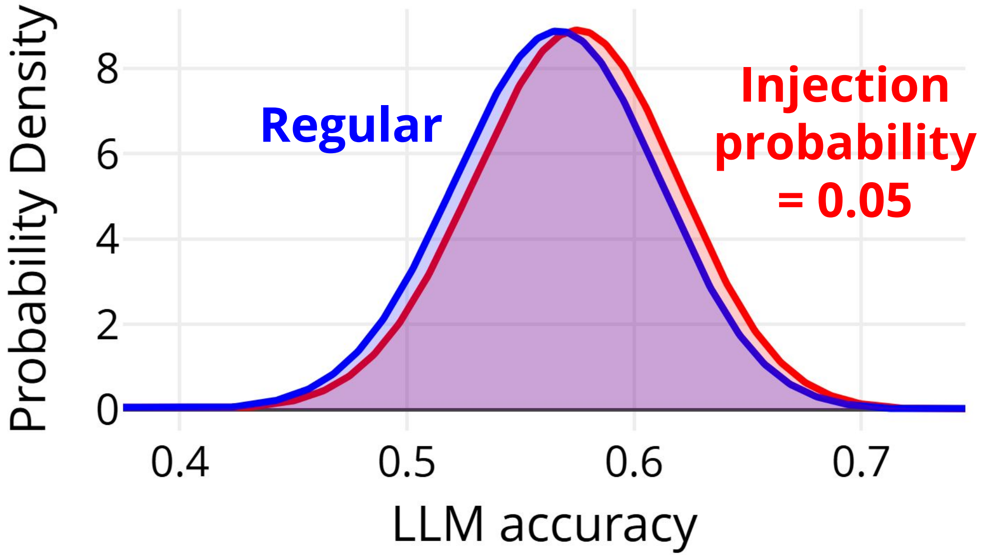
Doubt Injection – Results

- Adversarial questions: accuracy rise 26.1% → 26.7%
- Arithmetic reasoning: accuracy rise 56.7% → 57.5%
- Statistically insignificant (~58%, 55% probability it helps)
- Question-dependent



Doubt Injection – Results

- Adversarial questions: accuracy rise 26.1% → 26.7%
- Arithmetic reasoning: accuracy rise 56.7% → 57.5%
- Statistically insignificant (~58%, 55% probability it helps)
- Question-dependent
- String-dependent
- (“But” is best)



Conclusions

- DoLa on distribution with context helps CAD
- Additive CAD even better at resolving knowledge conflicts
- Doubt Injection shows limited potential