Mitigating Hallucinations in LLMs

Stephen Cowley, Magdalene College
Supervisor: Dr Marcus Tomalin

Large Language Model (LLM)

“Hello,
how are
you?”

Large Language Model (LLM)

Tokenisation

"Hello, how are you?"

$$\mathbf{w}_{<t}$$

$$= [w_1, w_2, \ldots, w_{t-1}]$$

Large Language Model (LLM)

Black box

Tokenisation

“Hello, how are you?” $\Rightarrow$ $\mathbf{w}_{<t}$ $\Rightarrow$ LLM

$= [w_1, w_2, \ldots, w_{t-1}]$

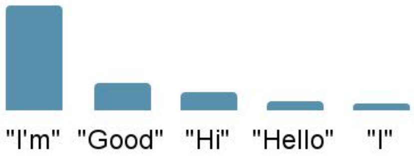
Large Language Model (LLM)

Tokenisation

Black box

“Hello, how are you?” $\Rightarrow$ $\mathbf{w}_{<t}$ $\Rightarrow$ LLM $\Rightarrow$ $\mathrm{P}(w_t \mid \mathbf{w}_{<t})$

$$= [w_1, w_2, \ldots, w_{t-1}]$$



Large Language Model (LLM)

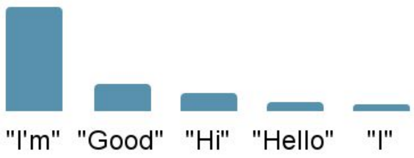
Tokenisation

Black box

Sample from distribution

“Hello, how are you?” → $\mathbf{w}_{<t}$ → LLM → $\mathrm{P}(w_t \mid \mathbf{w}_{<t})$ → “I’m”

$= [w_1, w_2, \ldots, w_{t-1}]$

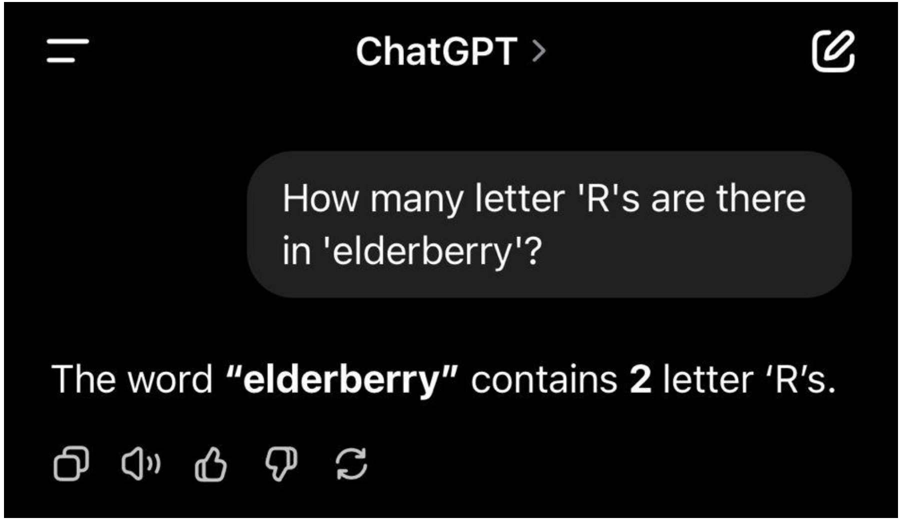


LLM Hallucination

- When “a model makes factual errors” – OpenAI

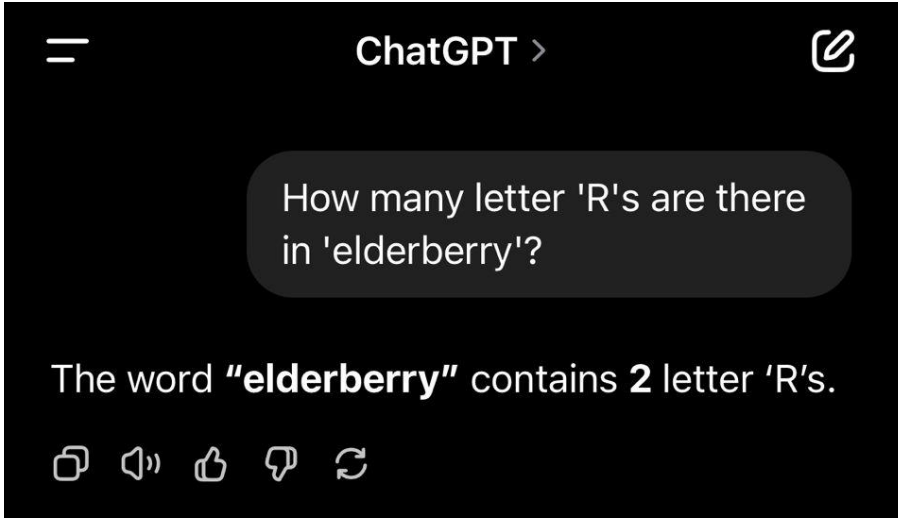
LLM Hallucination

- When “a model makes factual errors” – OpenAI



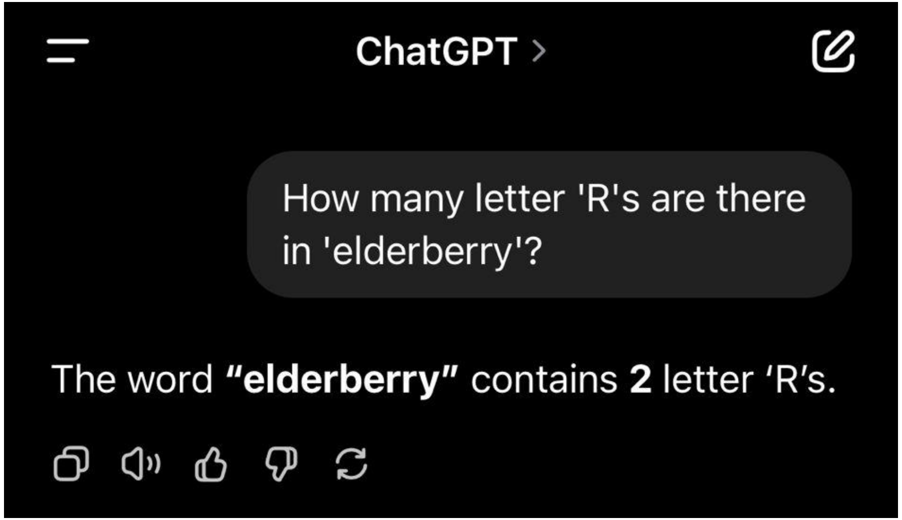
LLM Hallucination

- When “a model makes factual errors” – OpenAI



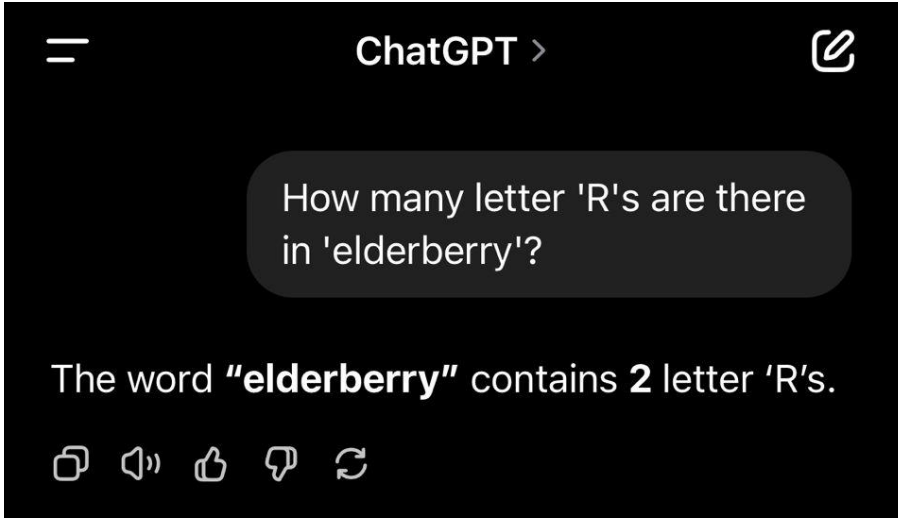
LLM Hallucination

- When “a model makes factual errors” – OpenAI
- Significant demand to improve LLM factual reliability



LLM Hallucination

- When “a model makes factual errors” – OpenAI
- Significant demand to improve LLM factual reliability
- 2 directions: decoding strategy; Chain-of-Thought modification




Forbes

BREAKING | BUSINESS

Lawyer Used ChatGPT In Court—And Cited Fake Cases. A Judge Is Considering Sanctions

By Molly Bohannon, Former Staff. Molly Bohannon ha...

Jun 08, 2023, 02:06pm EDT

Doubt Injection

- *Chain-of-Thought*
  is reasoning process

Doubt Injection

- *Chain-of-Thought* is reasoning process

Okay, so Alice and Bob are playing this game with a stack of n tokens. Alice goes first, and each turn a player can remove either 1 or 4 tokens. The person who takes the last token wins. I need to figure out how many positive integers n ≤ 2024 Bob can force a win by using a strategy, regardless of what Alice does. Hmm, this seems like a combinatorial game theory problem.

But I need to think through it step by step.

First, the grid is 5x5, so 25 cells in total. Each cell can have either a white chip, a black chip, or be empty. But with the constraints given, I need to make sure that in every row and every column, all the chips are the same color. That means if I have a row with some white chips, all the chips in that row must be white, and similarly for columns. The same applies to black chips.

But wait, n=1: Alice takes 1 token and wins. So Bob can't win when n=1.

But n=2: Alice can take 1 token, leaving 1 for Bob, who then takes it and wins. Or she can take 4, but n=2 is too small. So if n=2, Alice can win by taking 1 token. Wait, but actually, if n=2, Alice takes 1, leaves 1 for Bob, who takes it and wins. So Bob doesn't win. So n=2: Alice wins.

Doubt Injection

- *Chain-of-Thought* is reasoning process
- Try randomly inject e.g. “But” at new paragraph

Okay, so Alice and Bob are playing this game with a stack of n tokens. Alice goes first, and each turn a player can remove either 1 or 4 tokens. The person who takes the last token wins. I need to figure out how many positive integers $n \leq 2024$ Bob can force a win by using a strategy, regardless of what Alice does. Hmm, this seems like a combinatorial game theory problem.

But I need to think through it step by step.

First, the grid is 5x5, so 25 cells in total. Each cell can have either a white chip, a black chip, or be empty. But with the constraints given, I need to make sure that in every row and every column, all the chips are the same color. That means if I have a row with some white chips, all the chips in that row must be white, and similarly for columns. The same applies to black chips.

But wait, n=1: Alice takes 1 token and wins. So Bob can’t win when n=1.

But n=2: Alice can take 1 token, leaving 1 for Bob, who then takes it and wins. Or she can take 4, but n=2 is too small. So if n=2, Alice can win by taking 1 token. Wait, but actually, if n=2, Alice takes 1, leaves 1 for Bob, who takes it and wins. So Bob doesn’t win. So n=2: Alice wins.

Doubt Injection

- *Chain-of-Thought* is reasoning process
- Try randomly inject e.g. "But" at new paragraph
- Adversarial questions accuracy: 26.1% → **26.7%**
- Arithmetic reasoning accuracy: 56.7% → **57.5%**

Okay, so Alice and Bob are playing this game with a stack of n tokens. Alice goes first, and each turn a player can remove either 1 or 4 tokens. The person who takes the last token wins. I need to figure out how many positive integers n ≤ 2024 Bob can force a win by using a strategy, regardless of what Alice does. Hmm, this seems like a combinatorial game theory problem.

But I need to think through it step by step.

First, the grid is 5x5, so 25 cells in total. Each cell can have either a white chip, a black chip, or be empty. But with the constraints given, I need to make sure that in every row and every column, all the chips are the same color. That means if I have a row with some white chips, all the chips in that row must be white, and similarly for columns. The same applies to black chips.

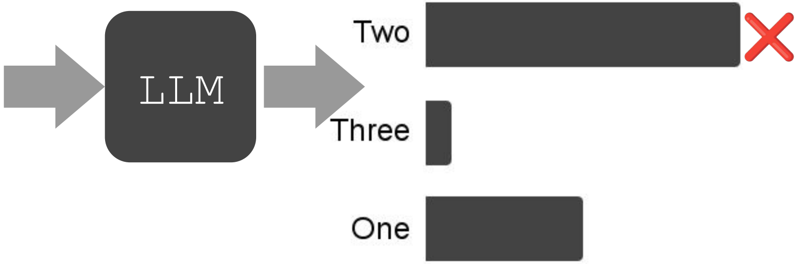
But wait, n=1: Alice takes 1 token and wins. So Bob can't win when n=1.

But n=2: Alice can take 1 token, leaving 1 for Bob, who then takes it and wins. Or she can take 4, but n=2 is too small. So if n=2, Alice can win by taking 1 token. Wait, but actually, if n=2, Alice takes 1, leaves 1 for Bob, who takes it and wins. So Bob doesn't win. So n=2: Alice wins.

Context-Aware Decoding (CAD)

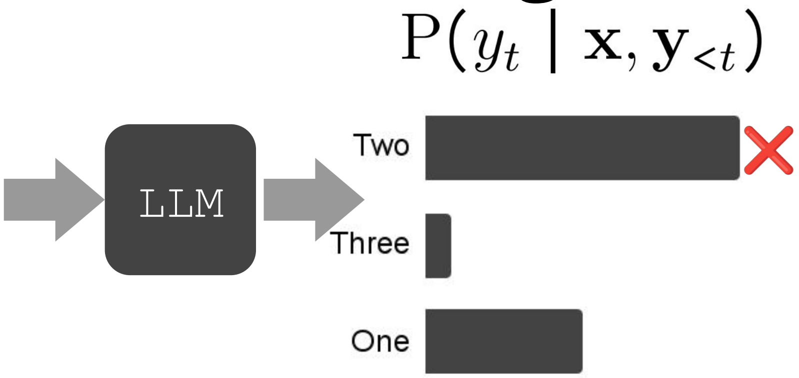
$$\mathrm{P}(y_t \mid \mathbf{x}, \mathbf{y}_{<t})$$

Query x
How many World Cups have Argentina won?

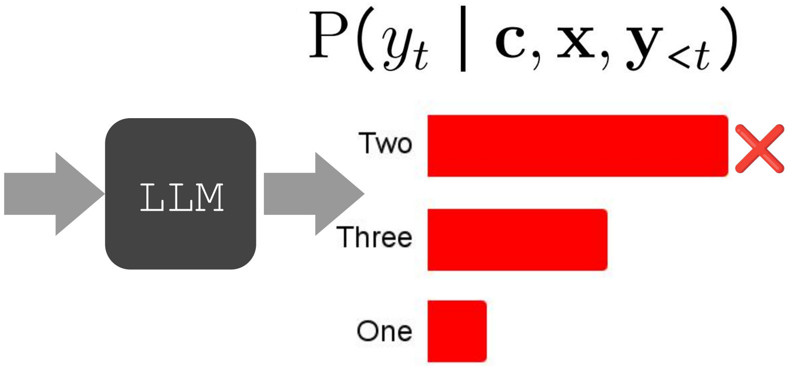


Context-Aware Decoding (CAD)

Query x
How many World Cups have Argentina won?



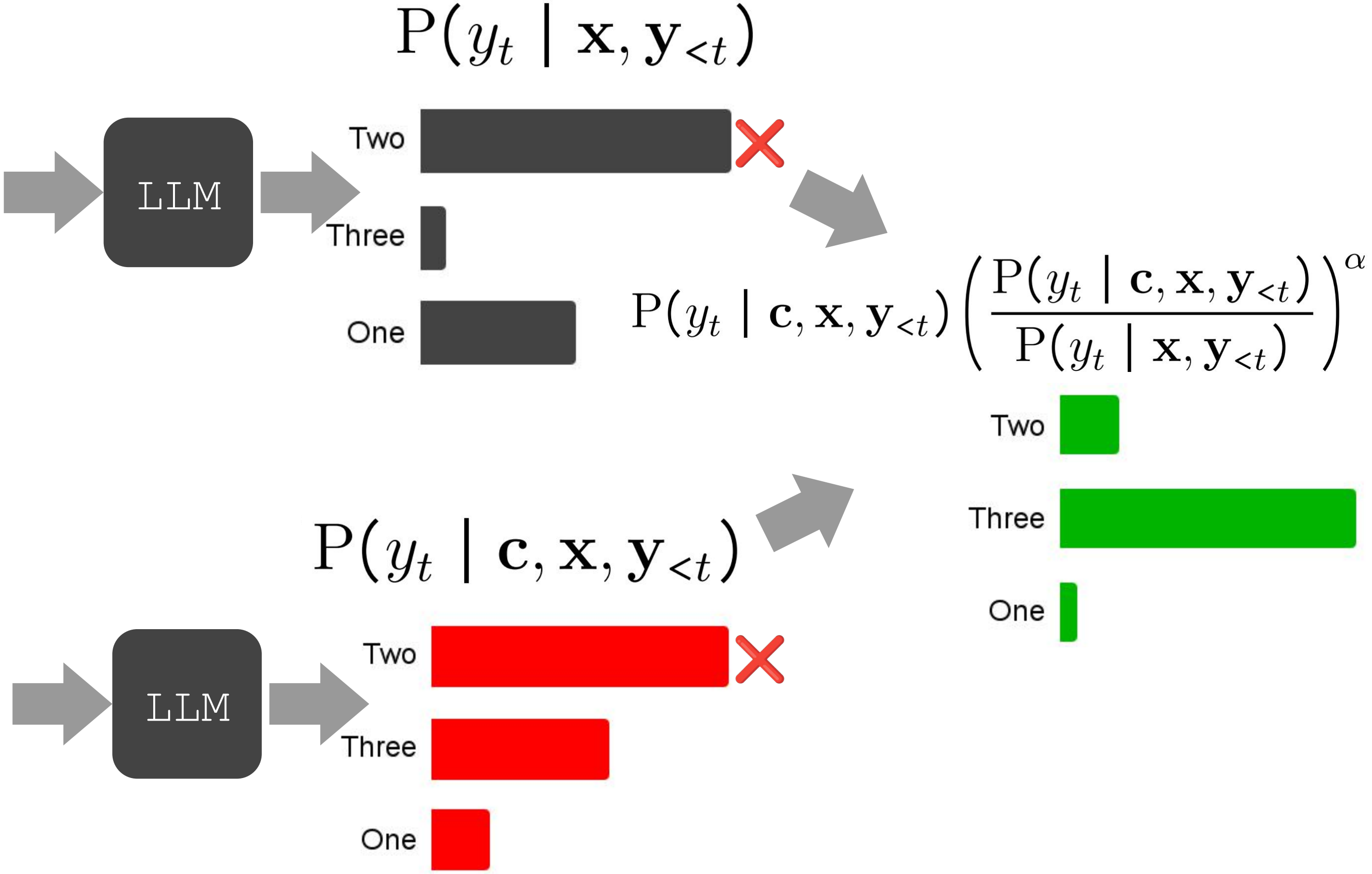
Context c
Argentina won World Cups in 1978, 1986 and 2022.
Query x
How many World Cups have Argentina won?



Context-Aware Decoding (CAD)

Query x
How many World Cups have Argentina won?

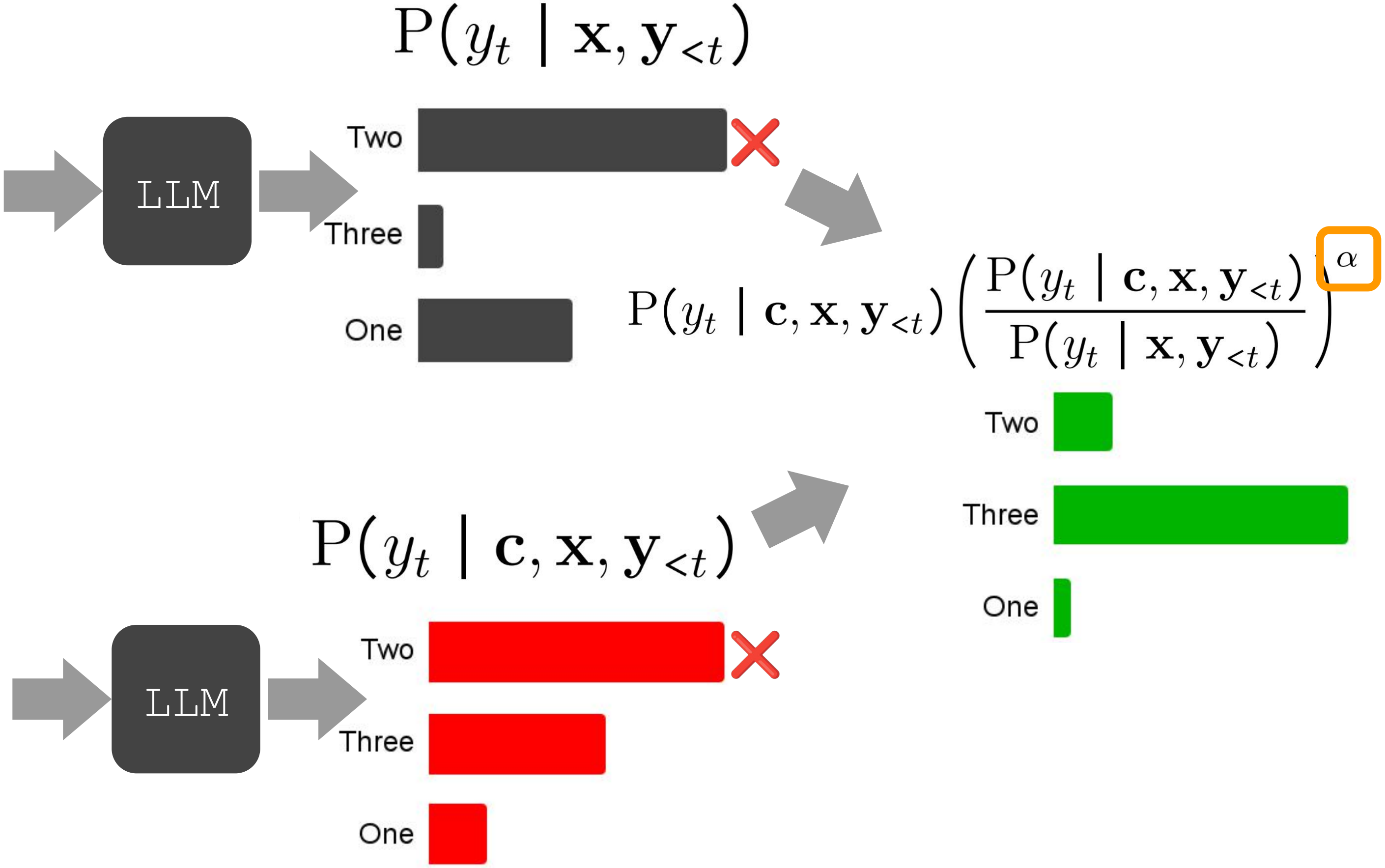
Context c
Argentina won World Cups in 1978, 1986 and 2022.
Query x
How many World Cups have Argentina won?



Context-Aware Decoding (CAD)

Query x
How many World Cups have Argentina won?

Context c
Argentina won World Cups in 1978, 1986 and 2022.
Query x
How many World Cups have Argentina won?



Results

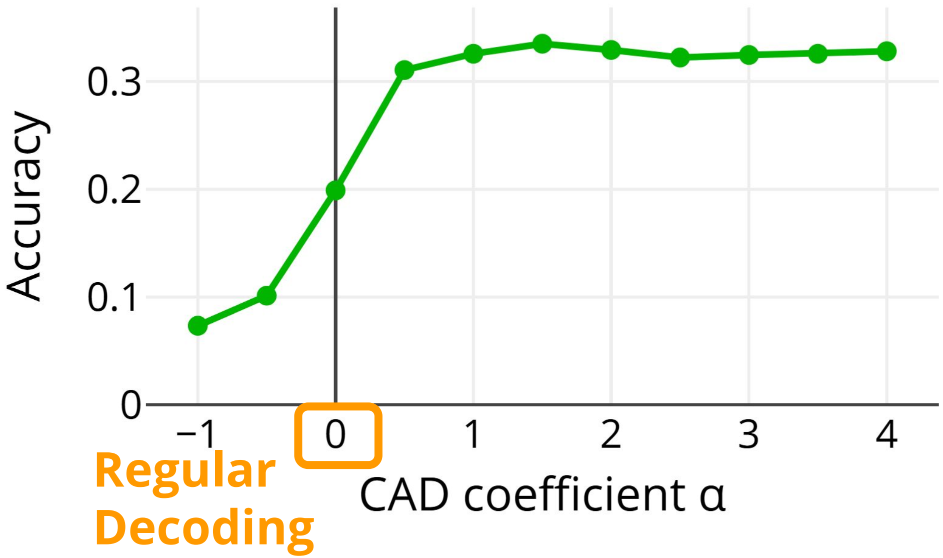
1. MemoTrap

Context c
Write a quote that ends in the word "early"
Query x
Better late than

Results

1. MemoTrap

Context c
Write a quote that ends in the word "early"
Query x
Better late than



Results

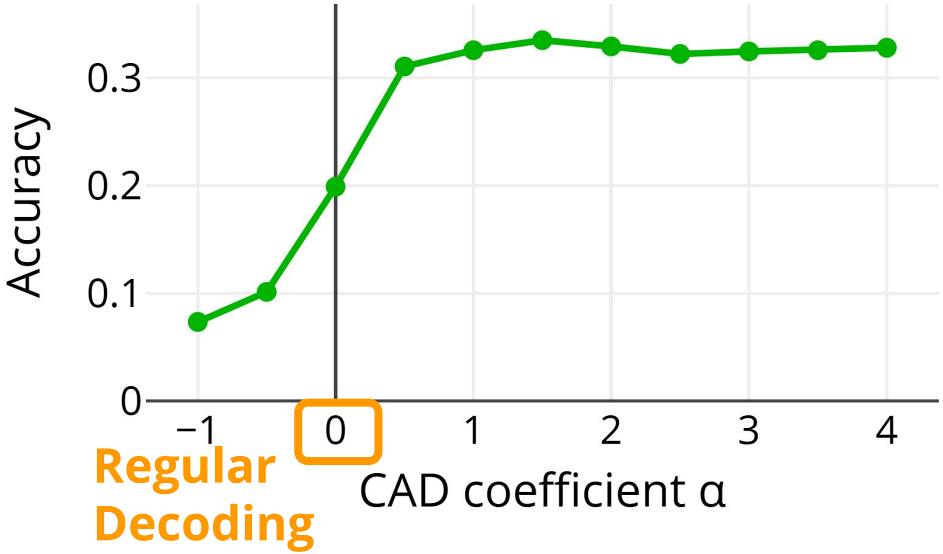
1. MemoTrap

Context c

Write a quote that ends in the word "early"

Query x

Better late than



2. Natural Questions

Context c

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

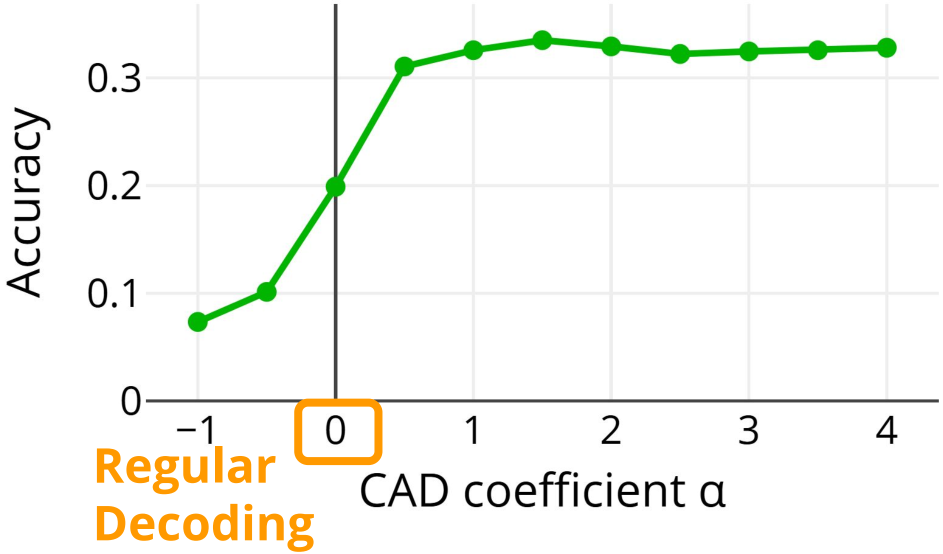
Query x

who holds the world record for the most world records?

Results

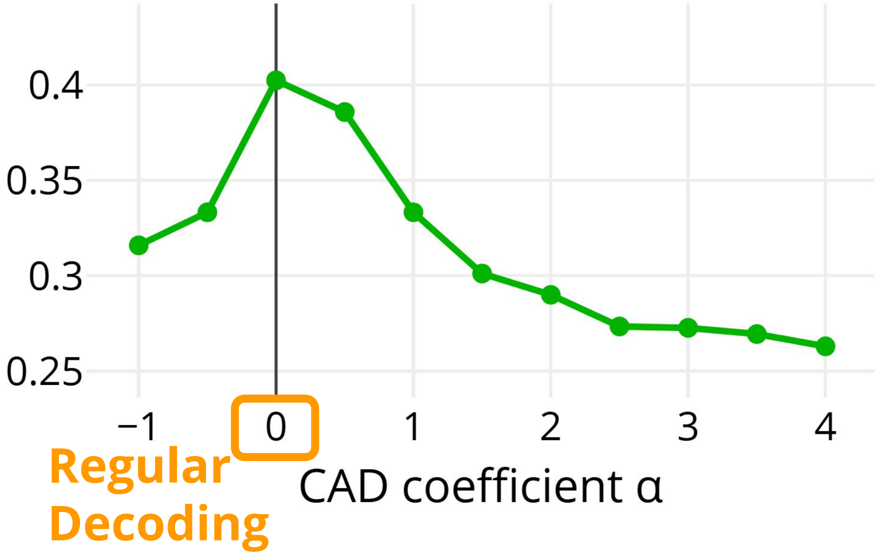
1. MemoTrap

Context c
Write a quote that ends in the word "early"
Query x
Better late than



2. Natural Questions

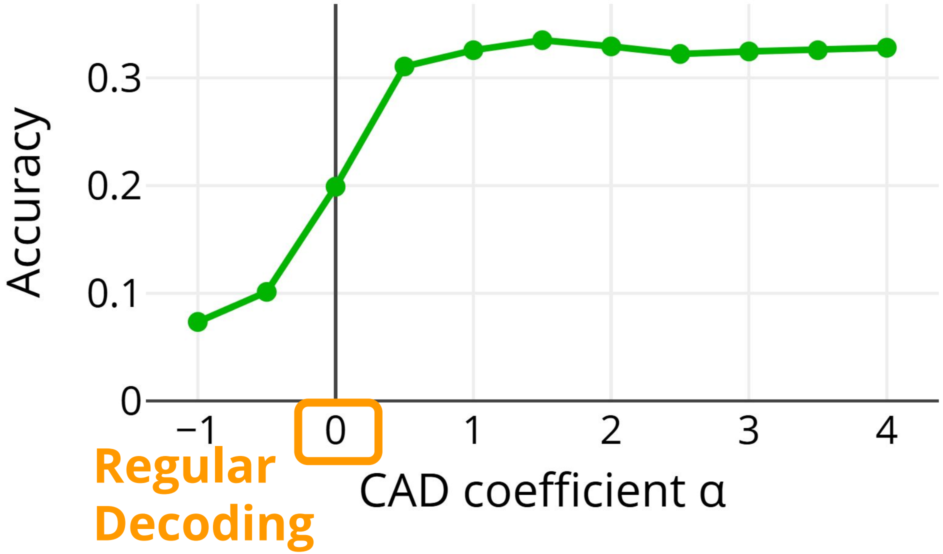
Context c
Ashrita Furman (born Keith Furman, September 16, 1954) is a Guinness World Records record-breaker…
Query x
who holds the world record for the most world records?



Results

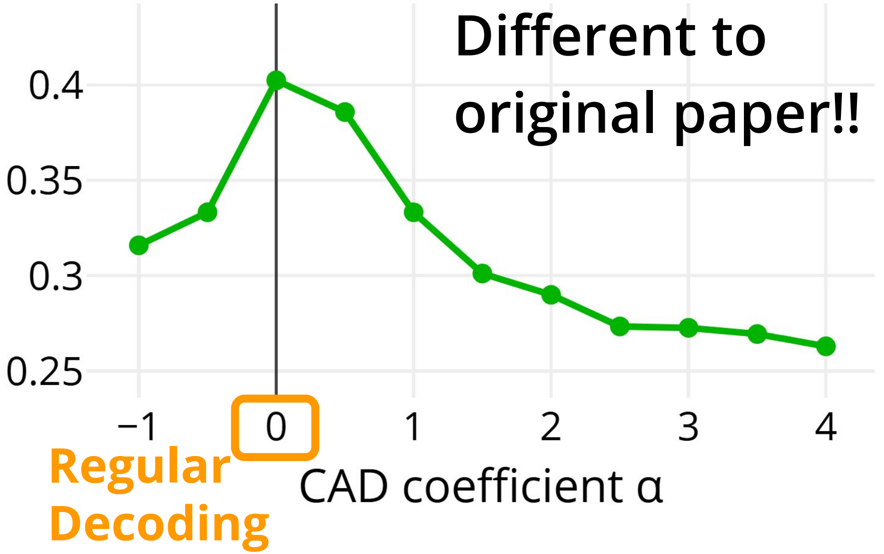
1. MemoTrap

Context c
Write a quote that ends in the word "early"
Query x
Better late than



2. Natural Questions

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



Different to original paper!!

Decoding by Contrasting Layers (DoLa)

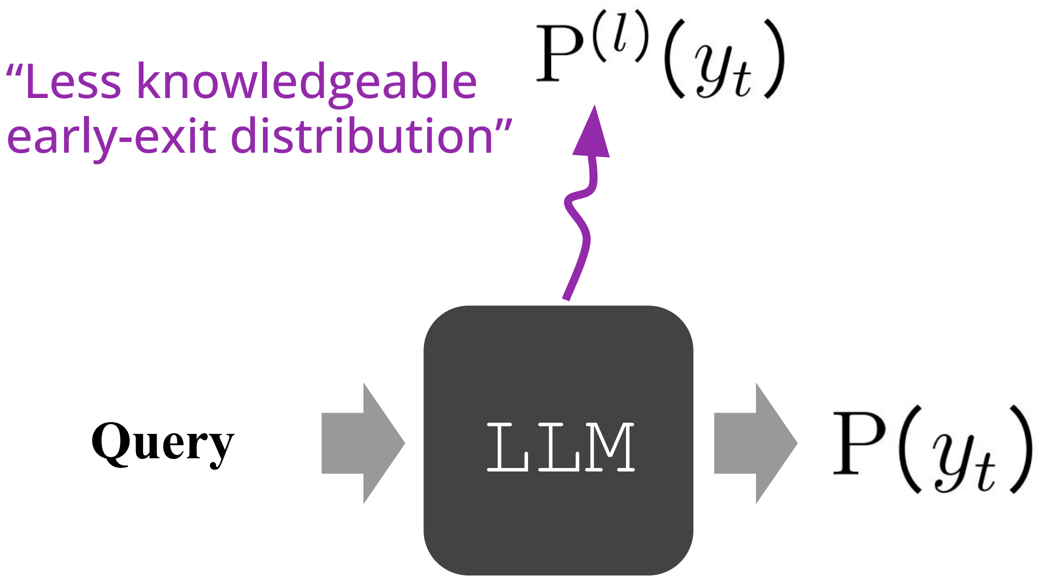
- Also contrasts distributions

Decoding by Contrasting Layers (DoLa)

- Also contrasts distributions

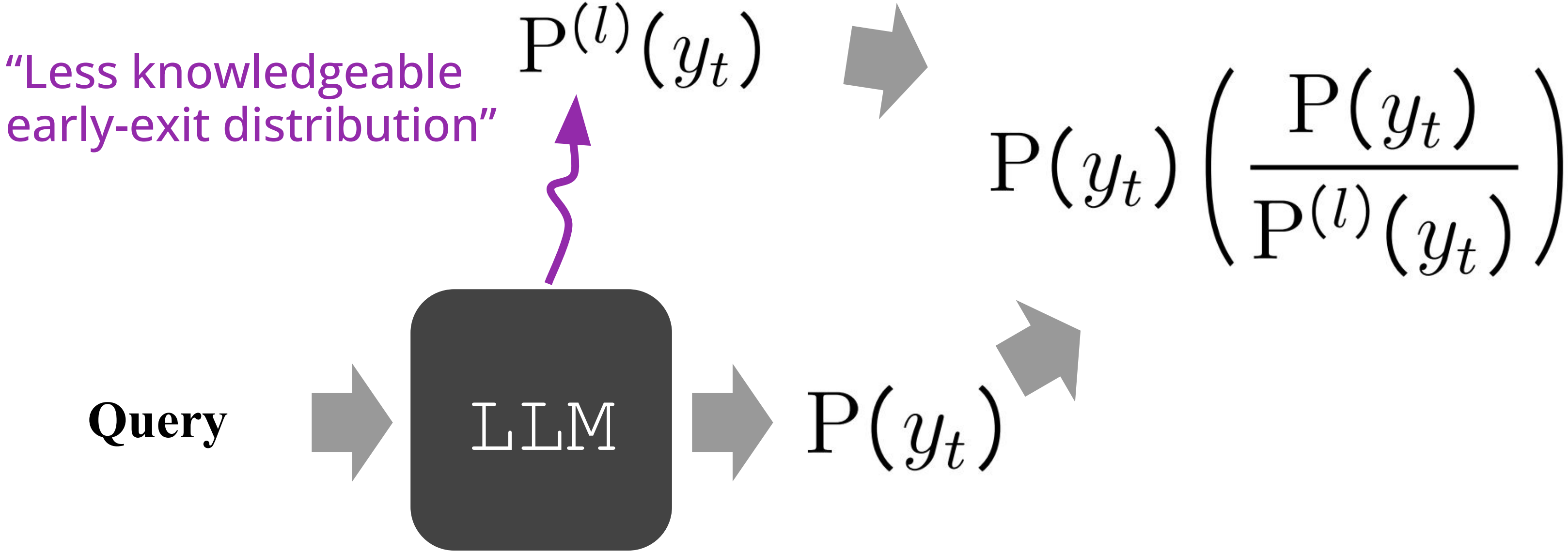
Decoding by Contrasting Layers (DoLa)

- Also contrasts distributions



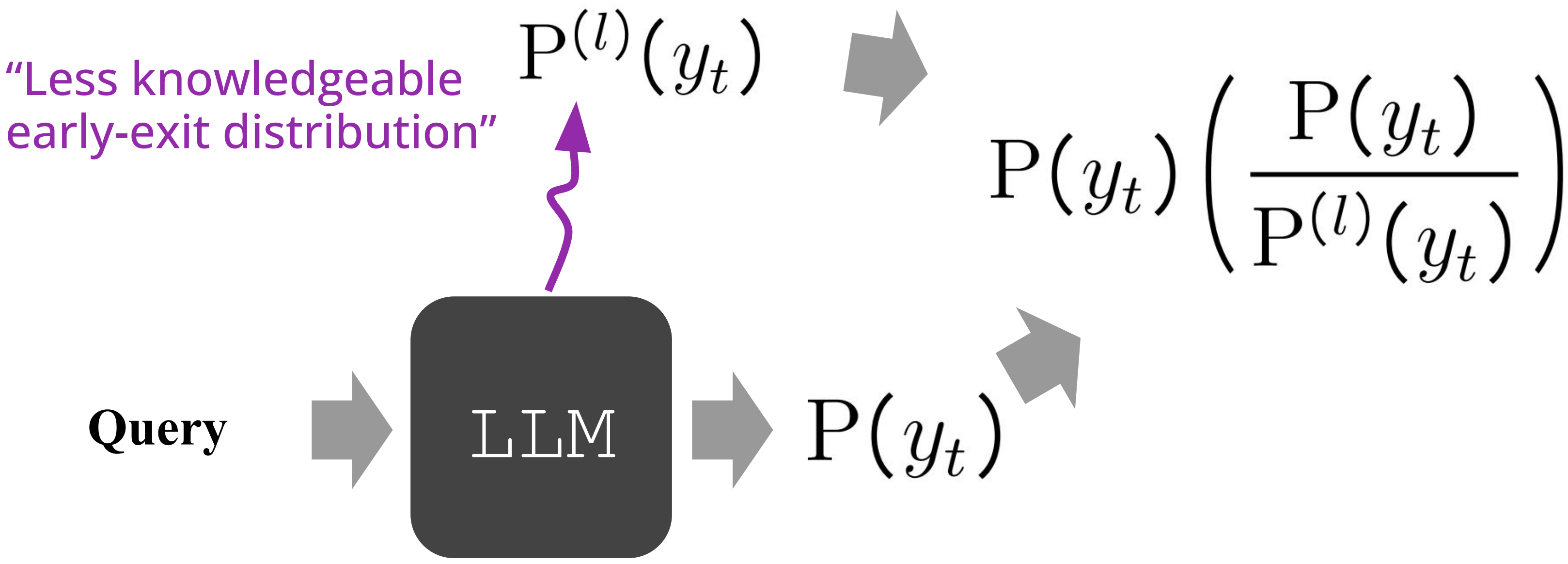
Decoding by Contrasting Layers (DoLa)

- Also contrasts distributions



Decoding by Contrasting Layers (DoLa)

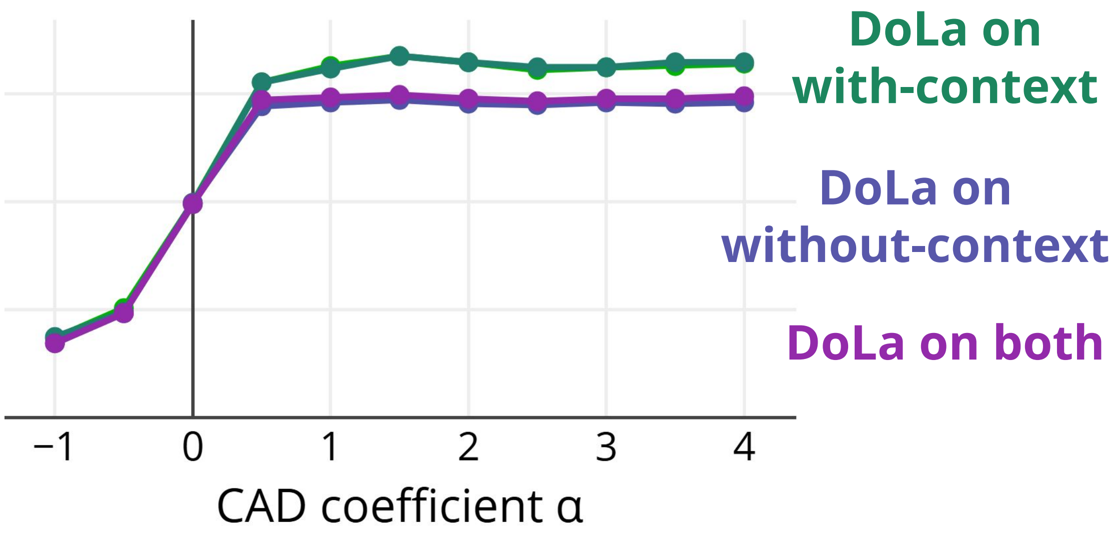
- Also contrasts distributions
- Increases factuality



Novel Combination

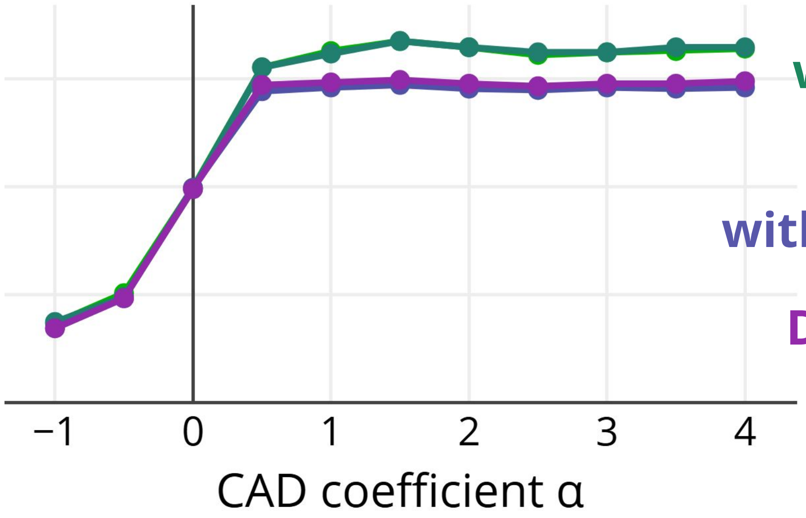
Novel Combination

1. MemoTrap



Novel Combination

1. MemoTrap



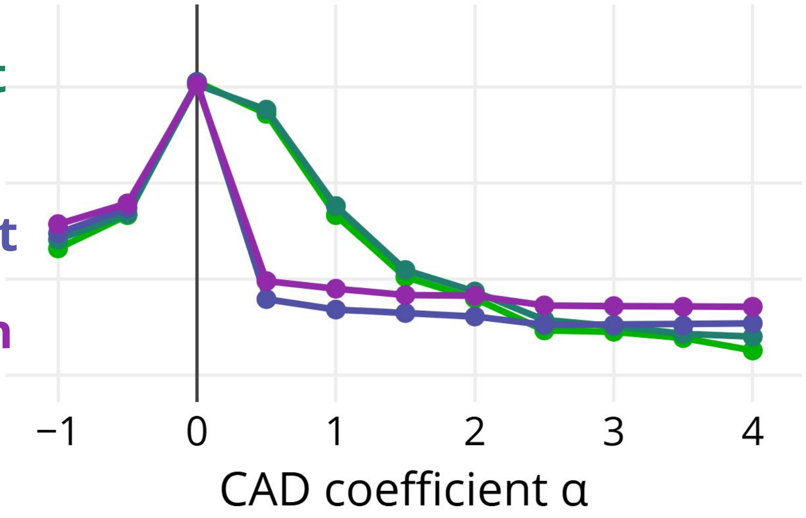
Normal

DoLa on with-context

DoLa on without-context

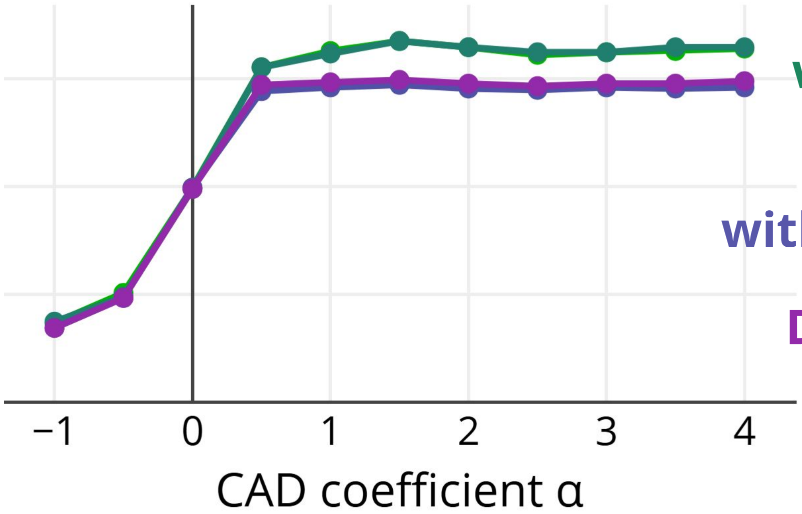
DoLa on both

2. Natural Questions



Novel Combination

1. MemoTrap



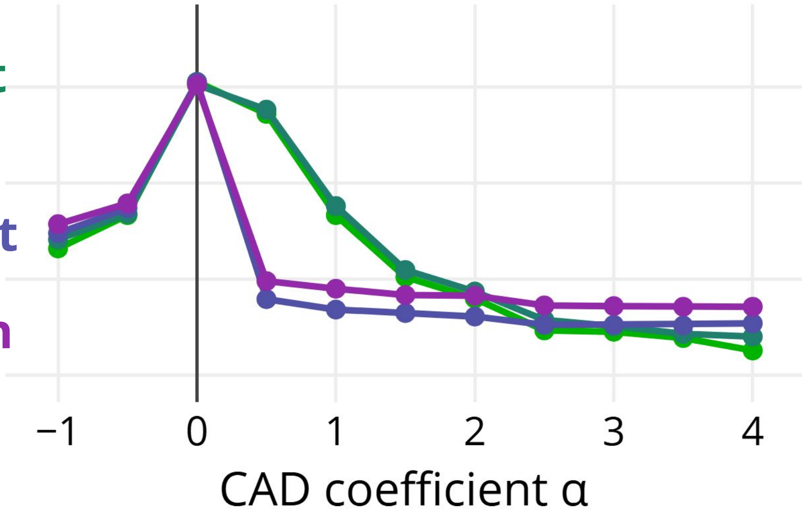
Normal

DoLa on with-context

DoLa on without-context

DoLa on both

2. Natural Questions



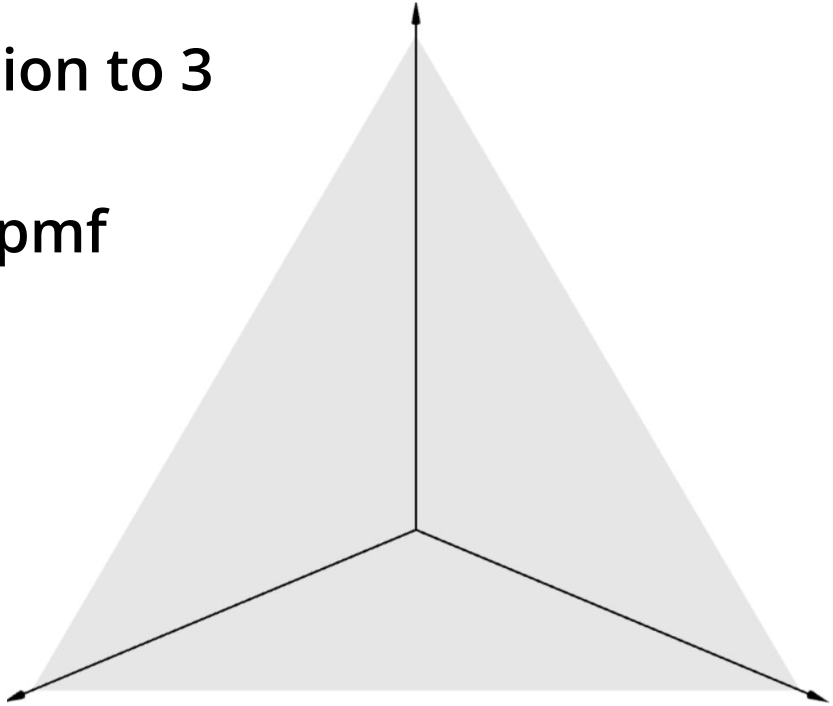
⇒ Performance improves when increasing factuality of distribution we're contrasting *in favour of*

Probability Simplex

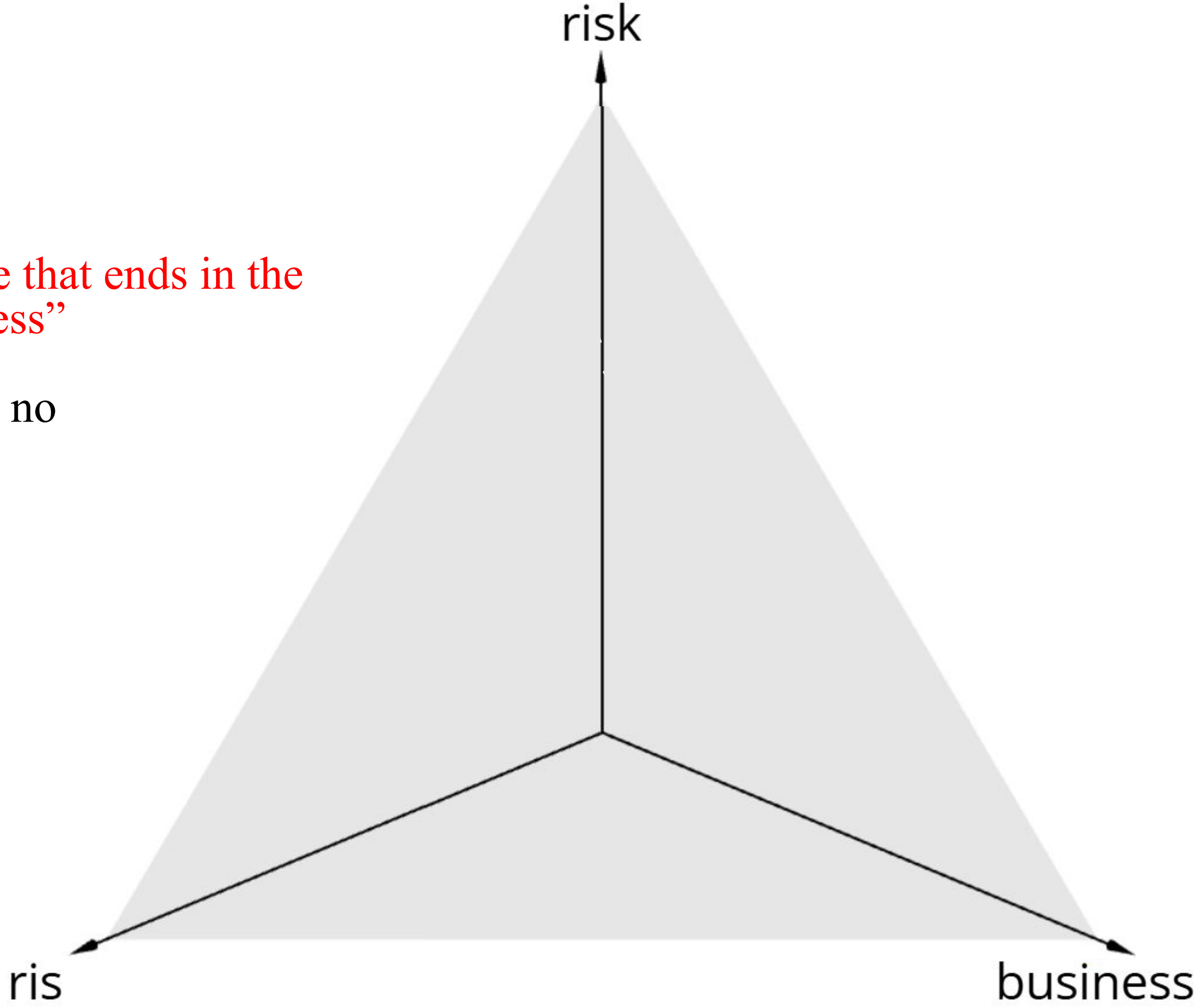
- Project 32,000 token distribution to 3 most relevant

Probability Simplex

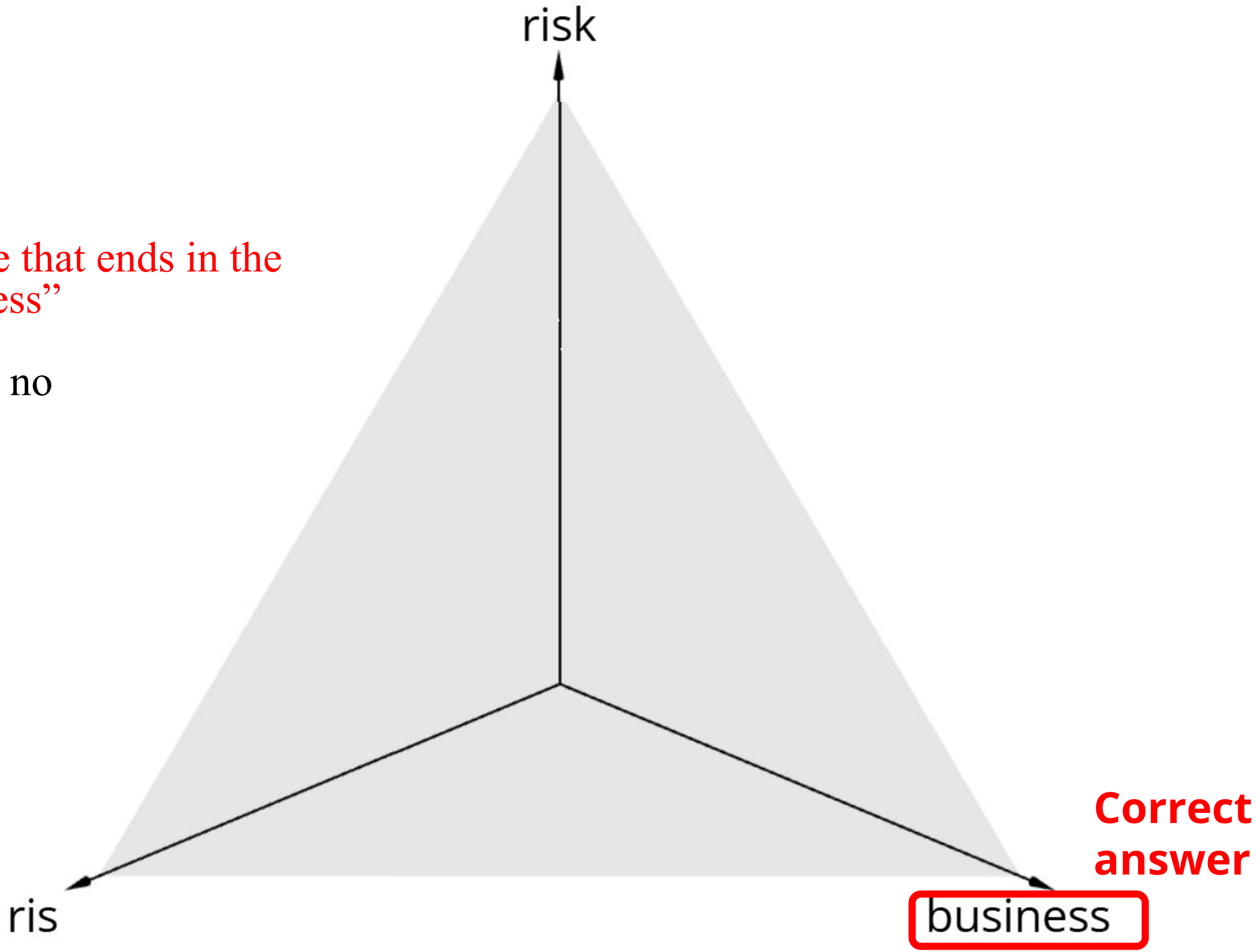
- Project 32,000 token distribution to 3 most relevant
- Triangle in 3D space satisfies pmf requirements



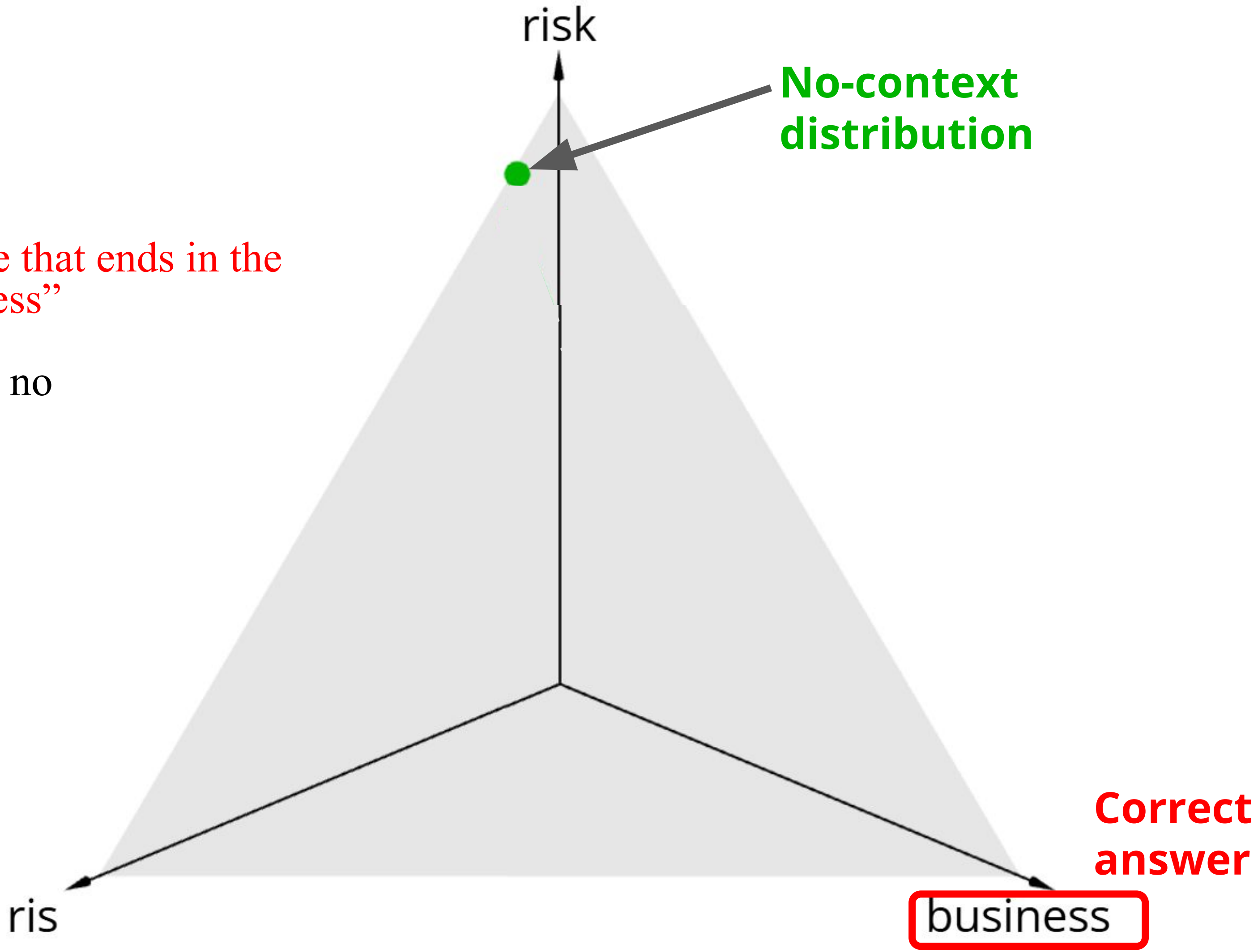
Context c
Write a quote that ends in the word "business"
Query x
Advisers run no



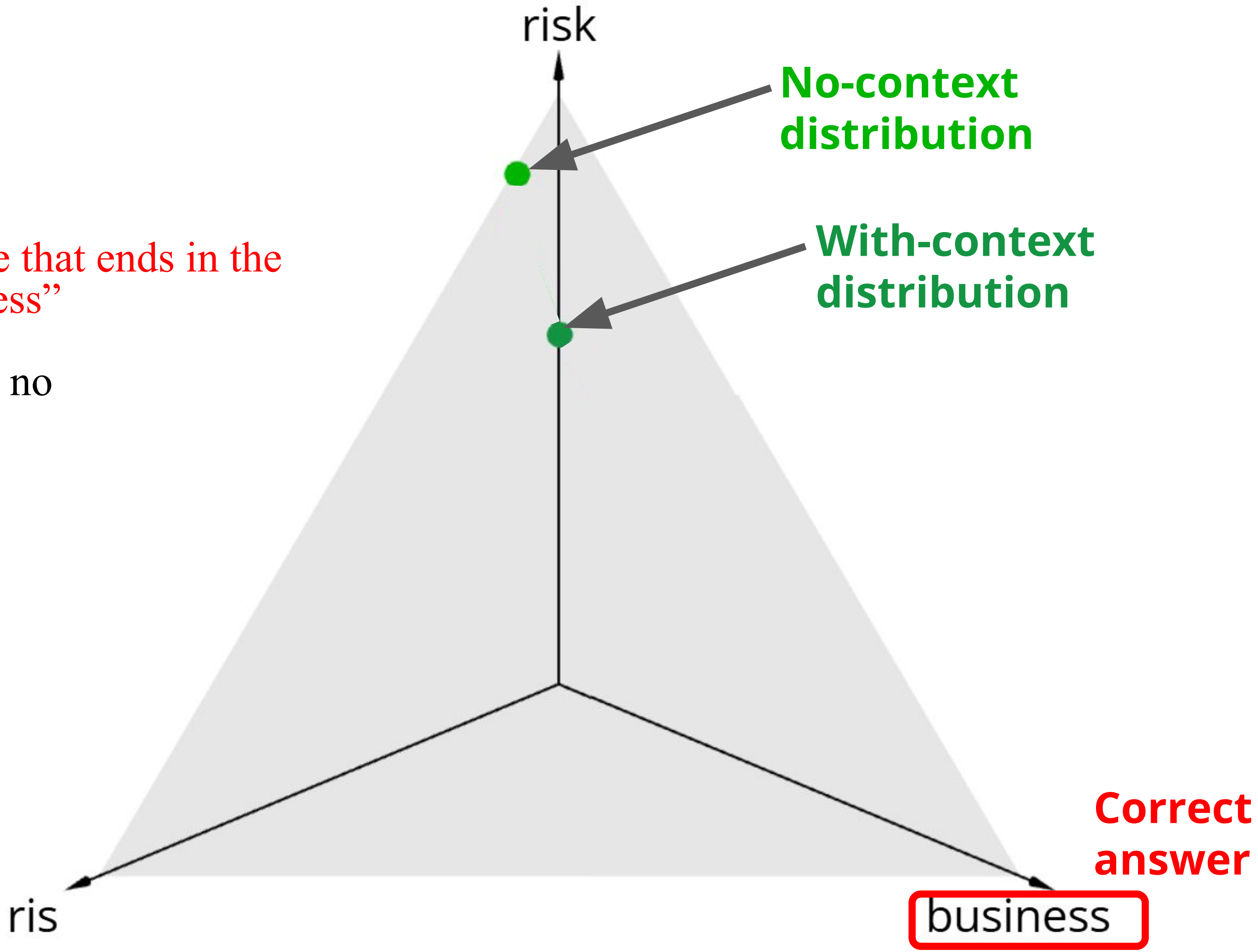
Context c
Write a quote that ends in the word “business”
Query x
Advisers run no



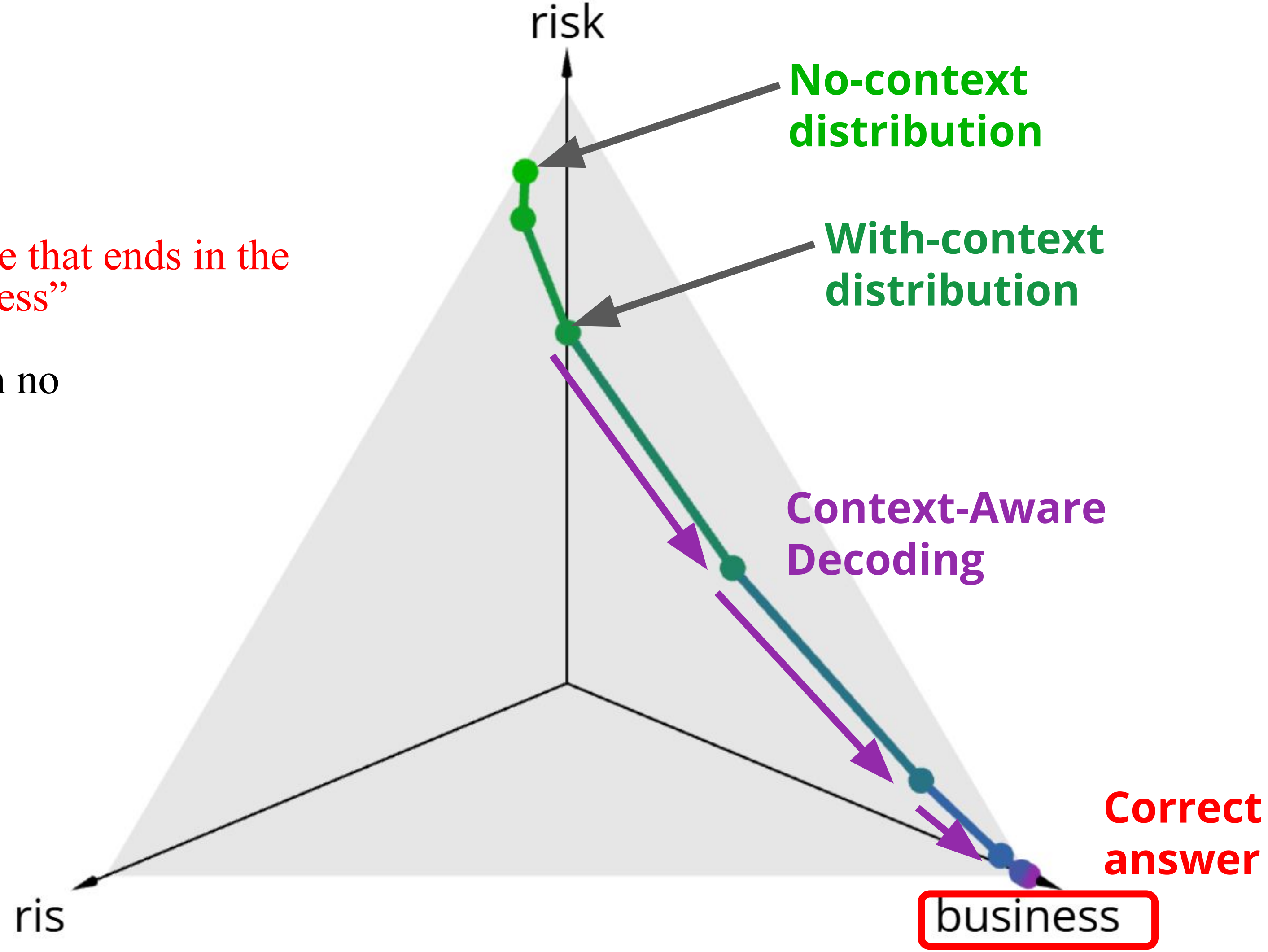
Context c
Write a quote that ends in the word "business"
Query x
Advisers run no

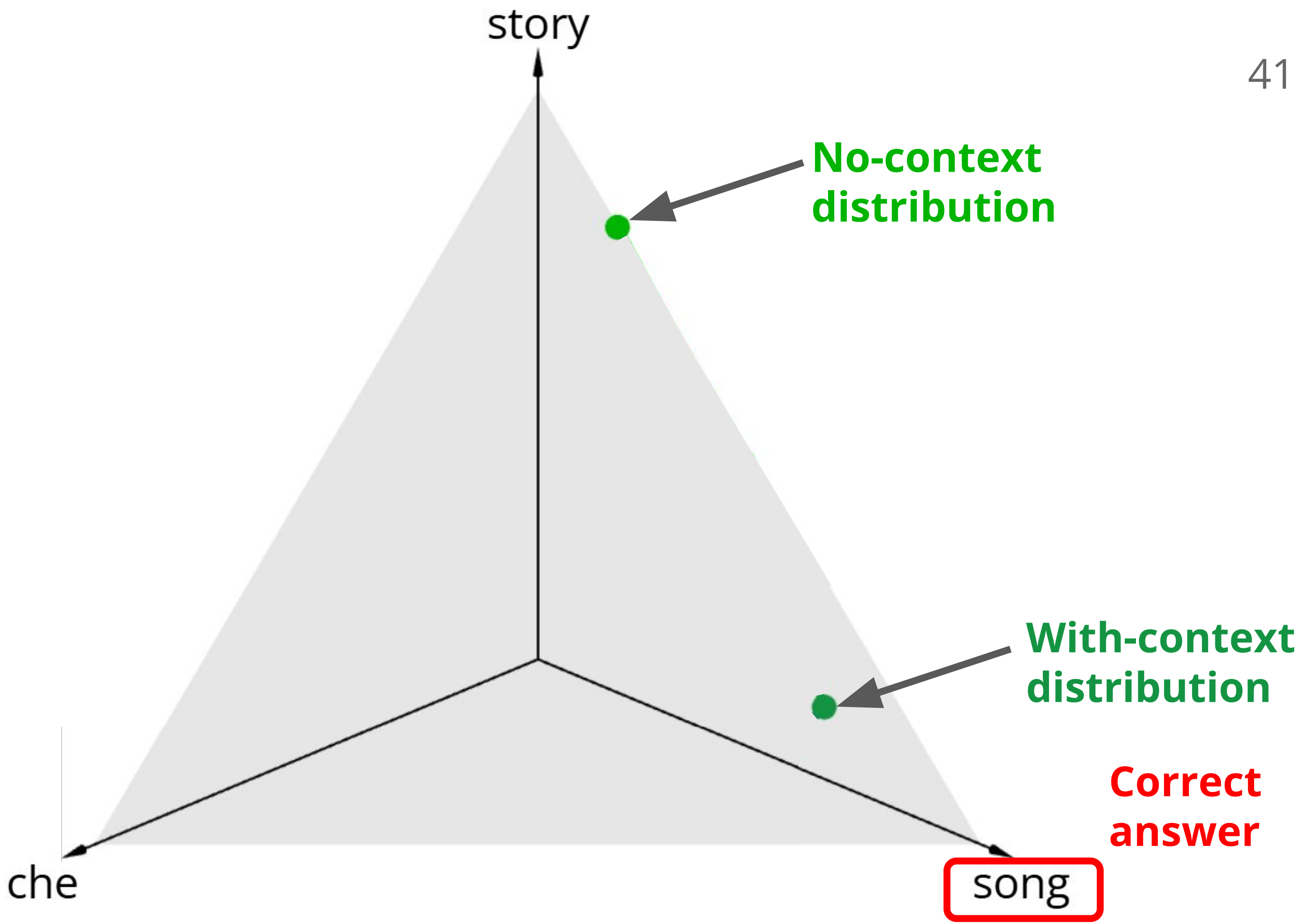


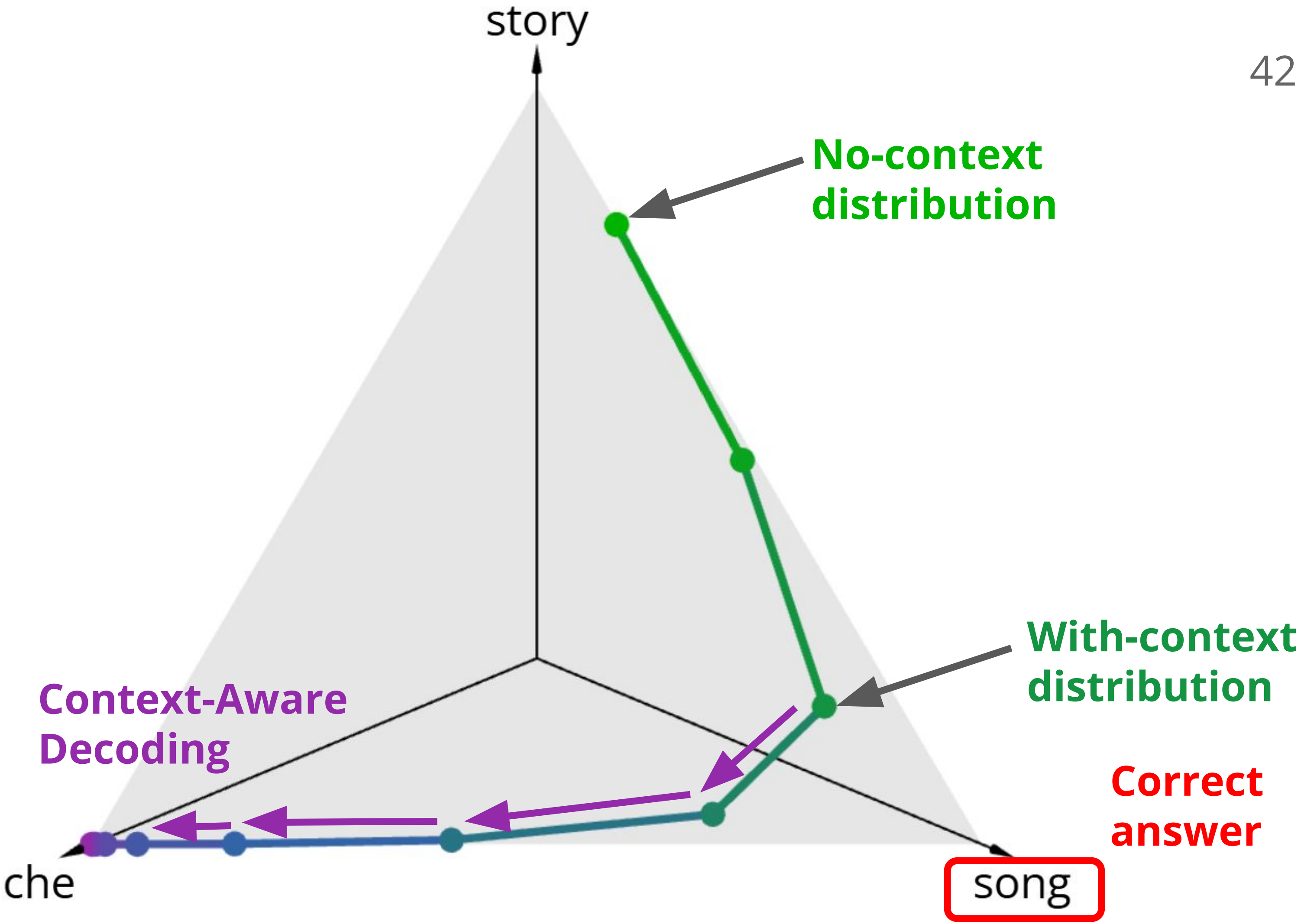
Context c
Write a quote that ends in the word “business”
Query x
Advisers run no

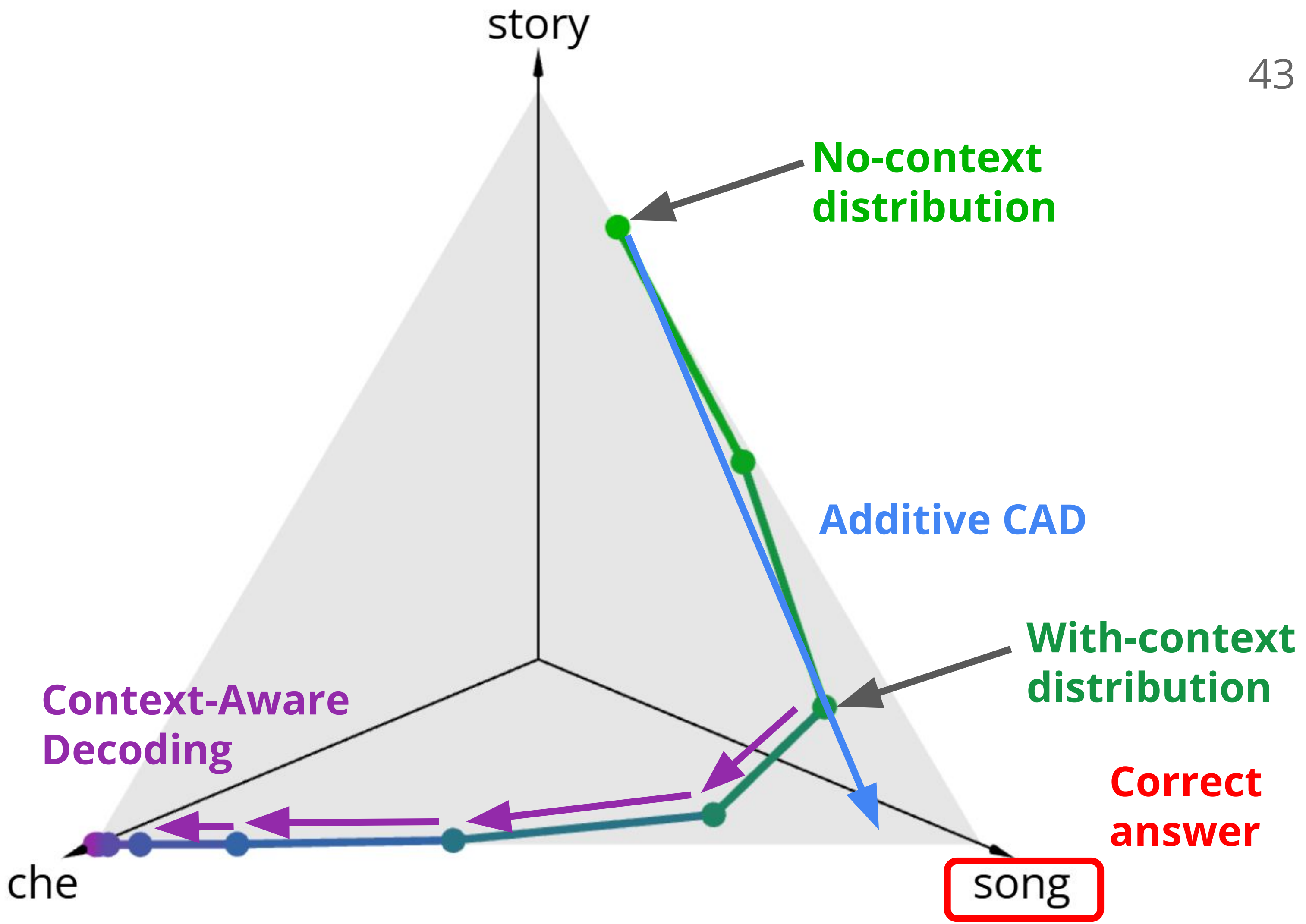


Context c
Write a quote that ends in the word "business"
Query x
Advisers run no




story
No-context
distribution
With-context
distribution
Correct
answer
che
song


story
No-context
distribution
With-context
distribution
Context-Aware
Decoding
Correct
answer
che
song


story
No-context
distribution
Additive CAD
With-context
distribution
Context-Aware
Decoding
Correct
answer
che
song

Additive Context-Aware Decoding

$$\mathrm{P}(y_t \mid \mathbf{x}, \mathbf{y}_{<t}) \ + \gamma \left(\mathrm{P}(y_t \mid \mathbf{c}, \mathbf{x}, \mathbf{y}_{<t}) - \mathrm{P}(y_t \mid \mathbf{x}, \mathbf{y}_{<t})\right)$$

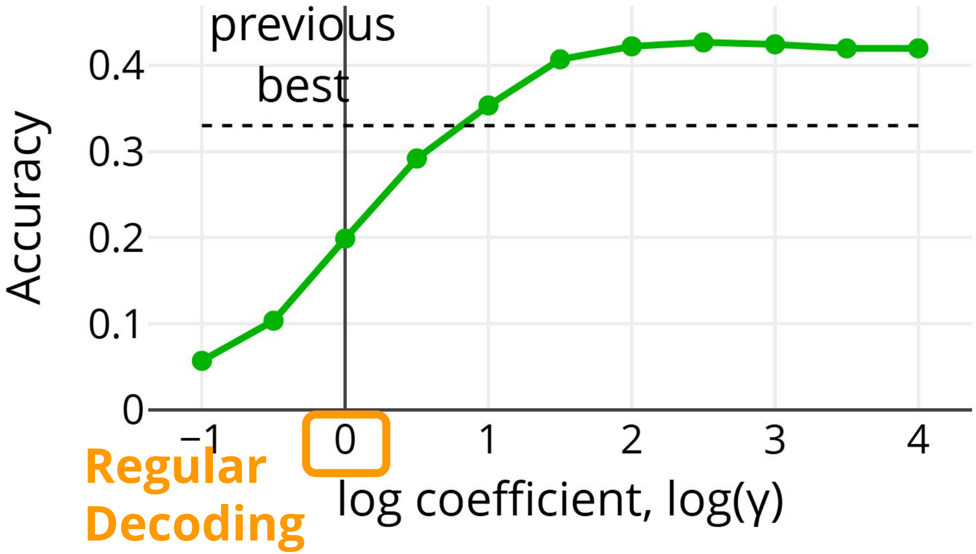
Additive Context-Aware Decoding

$$\mathrm{P}(y_t \mid \mathbf{x}, \mathbf{y}_{<t}) + \boxed{\gamma}\left(\mathrm{P}(y_t \mid \mathbf{c}, \mathbf{x}, \mathbf{y}_{<t}) - \mathrm{P}(y_t \mid \mathbf{x}, \mathbf{y}_{<t})\right)$$

Additive Context-Aware Decoding

$$\mathrm{P}(y_t \mid \mathbf{x}, \mathbf{y}_{<t}) + \boxed{\gamma}(\mathrm{P}(y_t \mid \mathbf{c}, \mathbf{x}, \mathbf{y}_{<t}) - \mathrm{P}(y_t \mid \mathbf{x}, \mathbf{y}_{<t}))$$

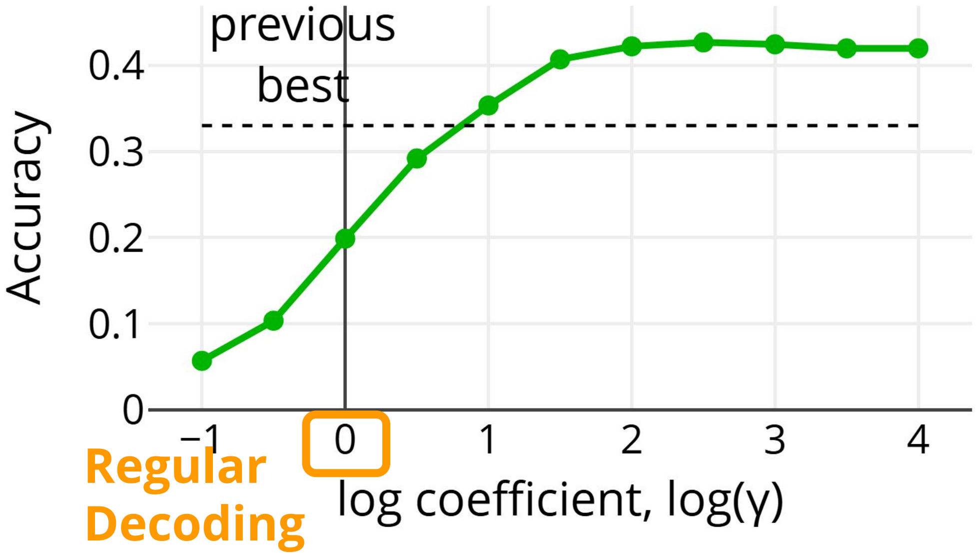
1. MemoTrap – even better



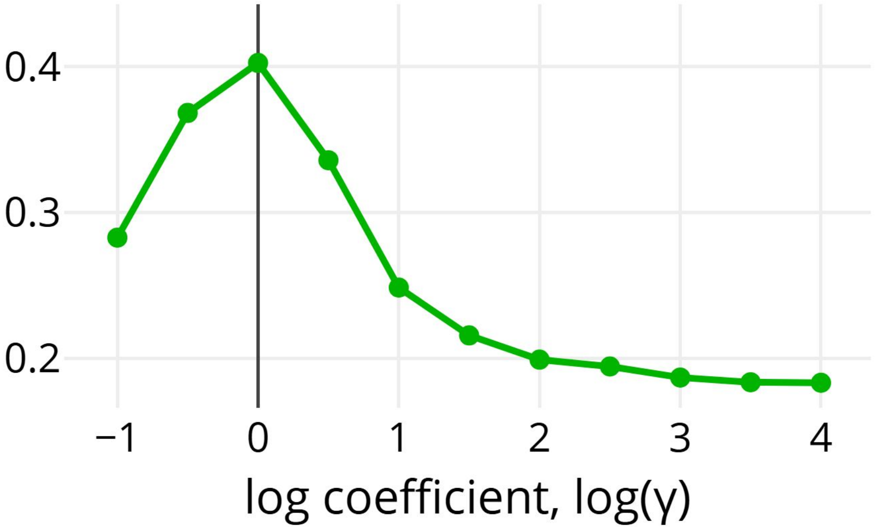
Additive Context-Aware Decoding

$$\mathrm{P}(y_t \mid \mathbf{x}, \mathbf{y}_{<t}) + \boxed{\gamma}(\mathrm{P}(y_t \mid \mathbf{c}, \mathbf{x}, \mathbf{y}_{<t}) - \mathrm{P}(y_t \mid \mathbf{x}, \mathbf{y}_{<t}))$$

1. MemoTrap – even better



2. Natural Questions – even worse



Conclusions

- DoLa on distribution *with context* helps CAD

Conclusions

- DoLa on distribution *with context* helps CAD
- Additive CAD *even better* at resolving knowledge conflicts

Conclusions

- DoLa on distribution *with context* helps CAD
- Additive CAD *even better* at resolving knowledge conflicts
- Doubt Injection shows some promise