Is my model perplexed for the right reason? Contrasting LLMs' Benchmark Behavior with Behavior Specified via Token-Level Perplexity

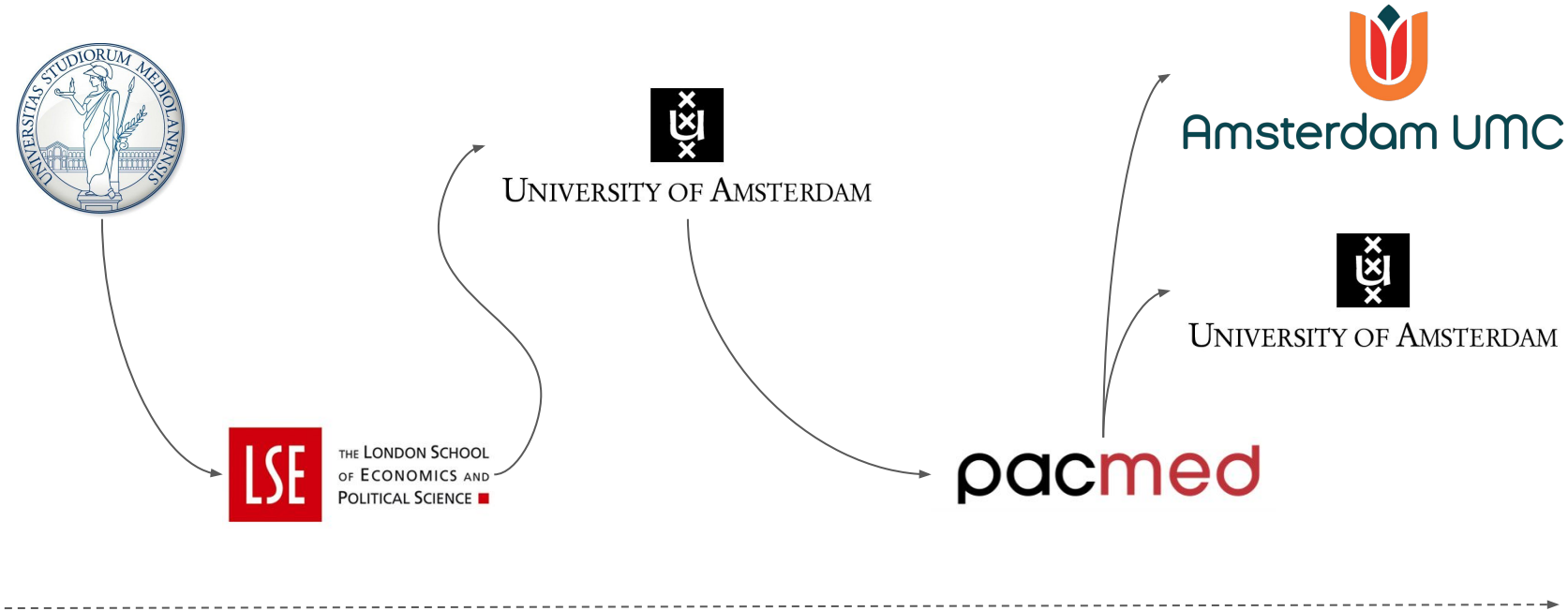
Giovanni Cinà

Amsterdam UMC, University of Amsterdam

✉ g.cina@amsterdamumc.nl

4-2-2026

My trajectory



My interests

In general, my work focuses on AI applications for healthcare.

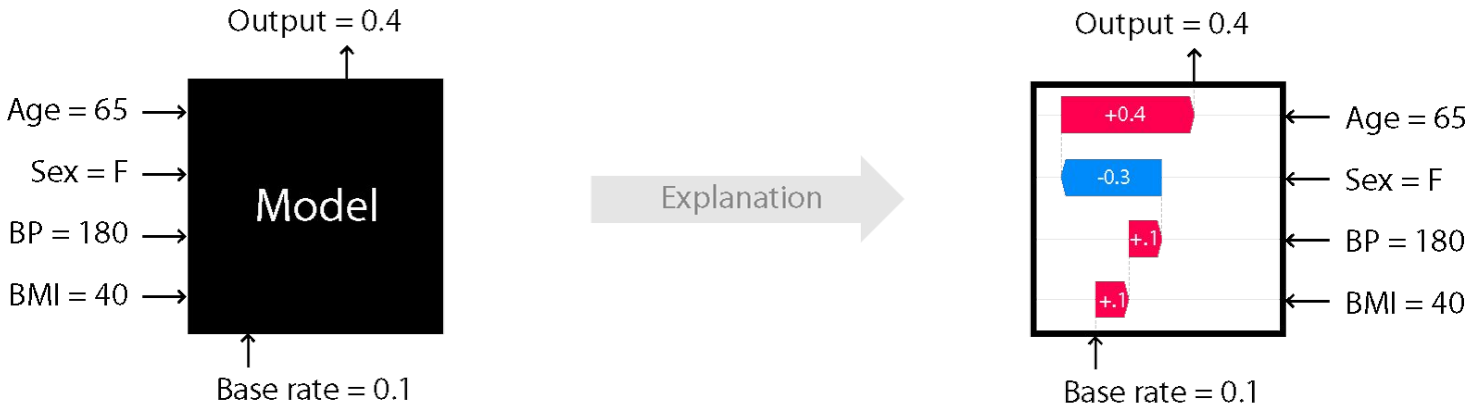
On the methodological side, I dabble in:

- Causal inference
- Out-Of-Distribution detection
- **Explainable AI**

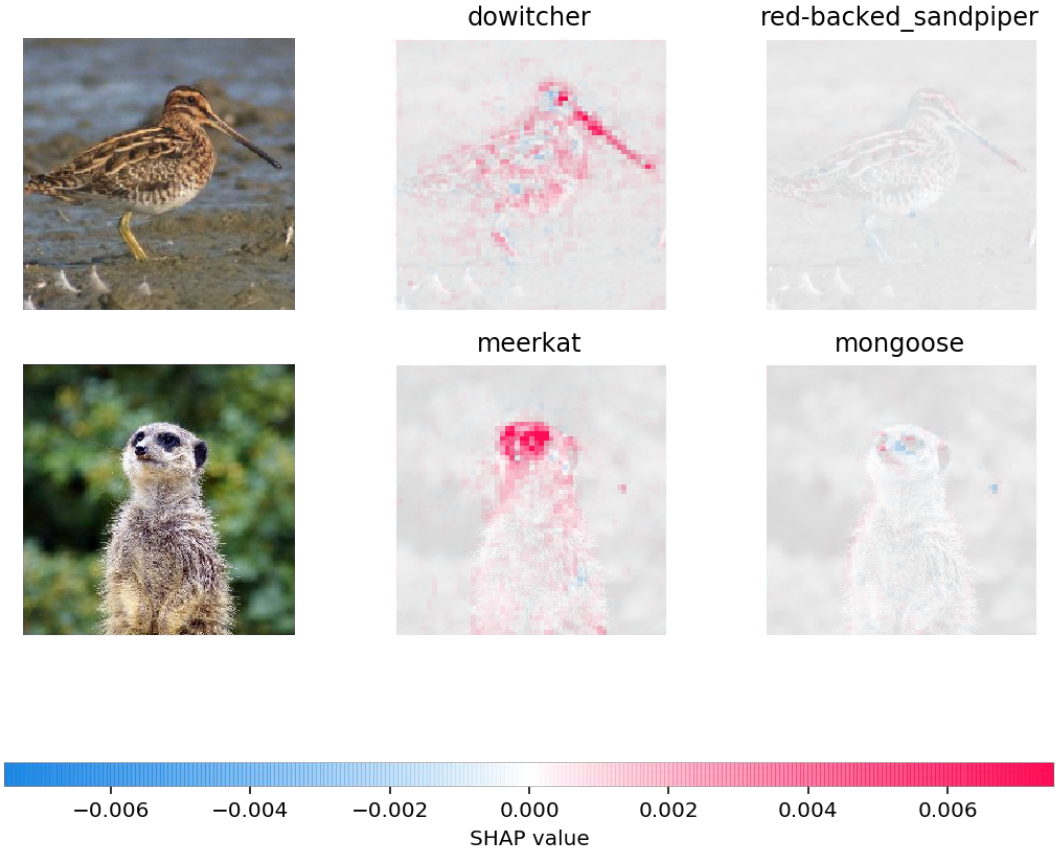
Agenda

1. Confirmation bias
2. Measuring linguistic abilities of LLMs without confirmation bias

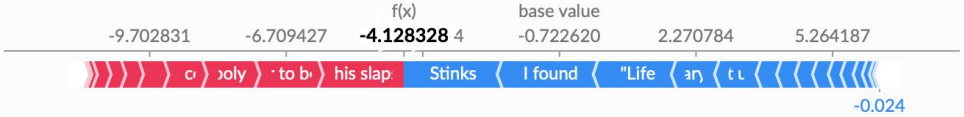
Example: classification on tabular data



Example: classification on image data



Example: classification on text



Homelessness (or Houselessness as George Carlin stated) has been an issue for years but never a plan to help those on the street that were once considered human who did everything from going to school, work, or vote for the matter. Most people think of the homeless as just a lost cause while worrying about things such as racism, the war on Iraq, pressuring kids to succeed, technology, the elections, inflation, or worrying if they'll be next to end up on the streets.

But what if you were given a bet to live on the streets for a month without the luxuries you once had from a home, the entertainment sets, a bathroom, pictures on the wall, a computer, and everything you once treasure to see what it's like to be homeless? That is Goddard Bolt's lesson.

Mel Brooks (who directs) who stars as Bolt plays a rich man who has everything in the world until deciding to make a bet with a sissy rival (Jeffery Tambor) to see if he can live in the streets for thirty days without the luxuries; if Bolt succeeds, he can do what he wants with a future project of making more buildings. The bet's on where Bolt is thrown on the street with a bracelet on his leg to monitor his every move where he can't step off the sidewalk. He's given the nickname Pepto by a vagrant after it's written on his forehead where Bolt meets other characters including a woman by the name of Molly (Lesley Ann Warren) an ex-dancer who got divorce before losing her home, and her pals Sailor (Howard Morris) and Fumes (Teddy Wilson) who are already used to the streets. They're survivors. Bolt isn't. He's not used to reaching mutual agreements like he once did when being rich where it's fight or flight, kill or be killed.

While the love connection between Molly and Bolt wasn't necessary to plot, I found "Life Stinks" to be one of Mel Brooks' observant films where prior to being a comedy, it shows a tender side compared to his slapstick work such as Blazing Saddles, Young Frankenstein, or Spaceballs for the matter, to show what it's like having something valuable before losing it the next day or on the other hand making a stupid bet like all rich people do when they don't know what to do with their money. Maybe they should give it to the homeless instead of using it like Monopoly money.

Or maybe this film will inspire you to help others.

What can go wrong: the way in which explanations are used and understood

Suppose you get an image and an explanation

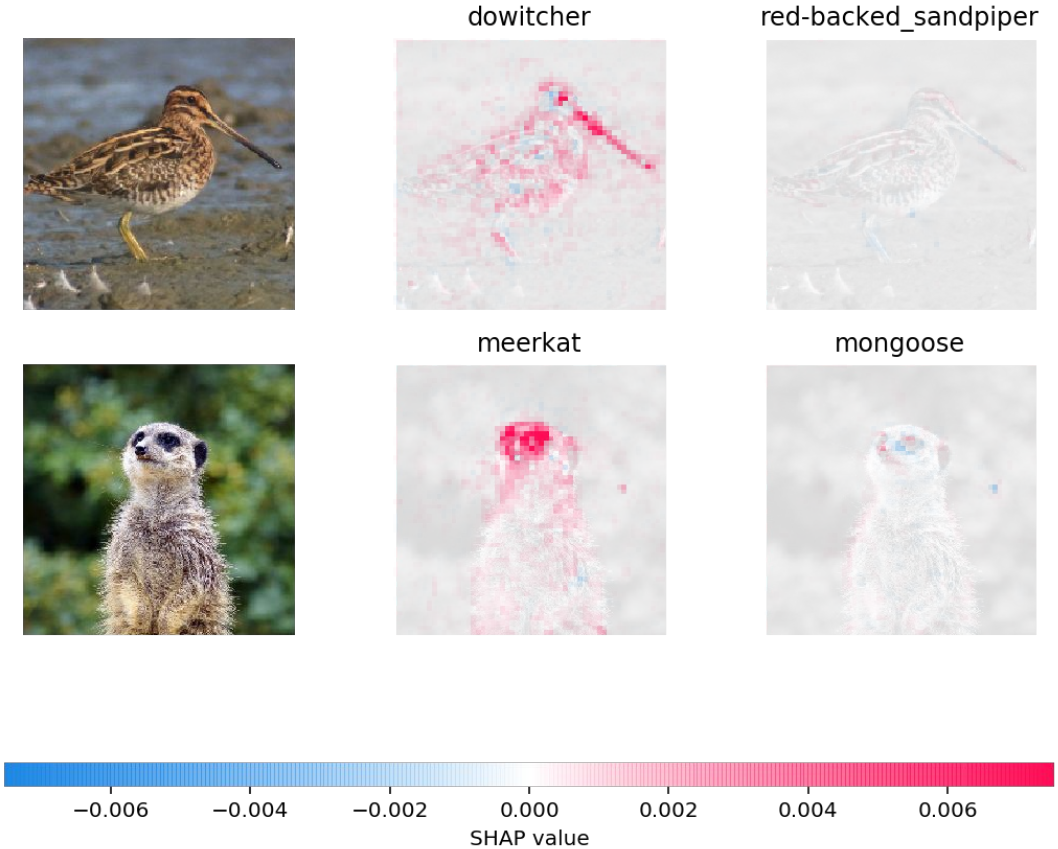
Is this convincing?

Why?

What can go wrong: the way in which explanations are used and understood

How do you know that the machine has a concept of ‘head’ that it is used to classify the meerkat?

Or ‘beak’ to classify the dowitcher?

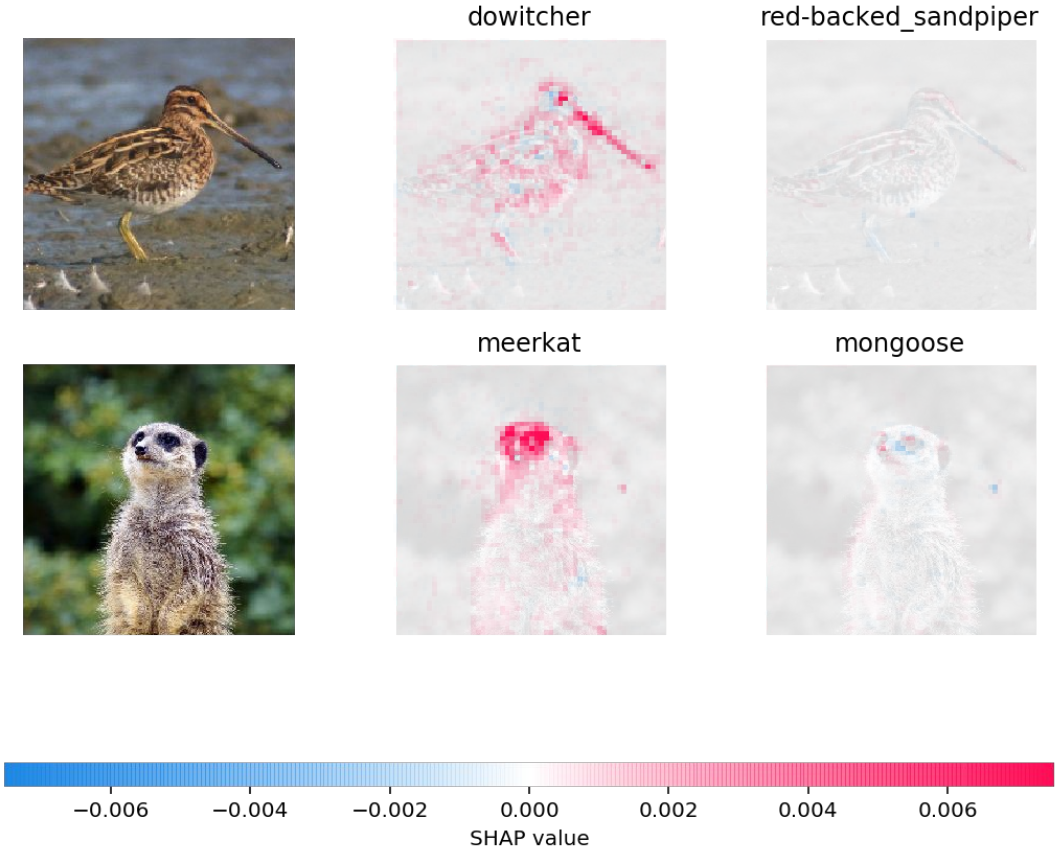


What can go wrong: confirmation bias

The fact that the cloud of pixels highlighted is sensible to us does not mean that it is highlighted for the right reason.

Confirmation bias

The tendency to believe explanations that confirm our belief/conviction.



Case study: linguistic abilities of LLMs

Exp. 1	*Is **S2** more specified than **S1**?* ✓
S1	The bag is on the chair. It is green
S2	The bag is on the chair. The chair is green.

Core idea: if an LLM can distinguish (classify) ambiguous from unambiguous sentences, it can tell the difference.

LLMs' abilities are assessed with benchmarks of minimal pairs

Various papers define benchmark to test specific linguistic abilities of LLMs.

Do Pre-Trained Language Models Detect and Understand Semantic Underspecification? Ask the DUST!

Frank Wildenburg, Michael Hanna, Sandro Pezzelle

BLiMP: The Benchmark of Linguistic Minimal Pairs for English

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, Samuel R. Bowman

We believe performance on benchmark is a too indirect way to test behaviour

Our research agenda:

- find a way to formally specify a behavior of interest
- find a metric to measure compliance with the behavior
- side-step confirmation bias

Our approach for LLMs' linguistic abilities

- find a way to formally specify a behavior of interest -> express behavior in terms of token-level perplexity
- find a metric to measure compliance with the behavior -> prompt LLMs and measure token-level perplexity metrics
- side-step confirmation bias -> make proper conclusions about what LLMs do and do not

Example of two prompts

- **correct:** *"This is an ambiguous sentence: 'Andrei approached the person **with** a green chair'. This is its unambiguous counterpart: 'Andrei approached the person **who had** a green chair'."*

- **incorrect:** *"This is an ambiguous sentence: 'Andrei approached the person **who had** a green chair'. This is its unambiguous counterpart: 'Andrei approached the person **with** a green chair'."*

Pivotal tokens

The tokens in **bold** are those that render a sentence (un)ambiguous.

- **correct:** *"This is an ambiguous sentence: 'Andrei approached the person **with** a green chair'. This is its unambiguous counterpart: 'Andrei approached the person **who had** a green chair'."*
- **incorrect:** *"This is an ambiguous sentence: 'Andrei approached the person **who had** a green chair'. This is its unambiguous counterpart: 'Andrei approached the person **with** a green chair'."*

Our definition of the desired behavior

If the model understands the difference between ambiguous and unambiguous, it should be *more perplexed at the pivotal tokens in bold in the second/incorrect prompt.*

- **correct:** *"This is an ambiguous sentence: 'Andrei approached the person* ***with*** *a green chair'. This is its unambiguous counterpart: 'Andrei approached the person* ***who had*** *a green chair'."*
- **incorrect:** *"This is an ambiguous sentence: 'Andrei approached the person* ***who had*** *a green chair'. This is its unambiguous counterpart: 'Andrei approached the person* ***with*** *a green chair'."*

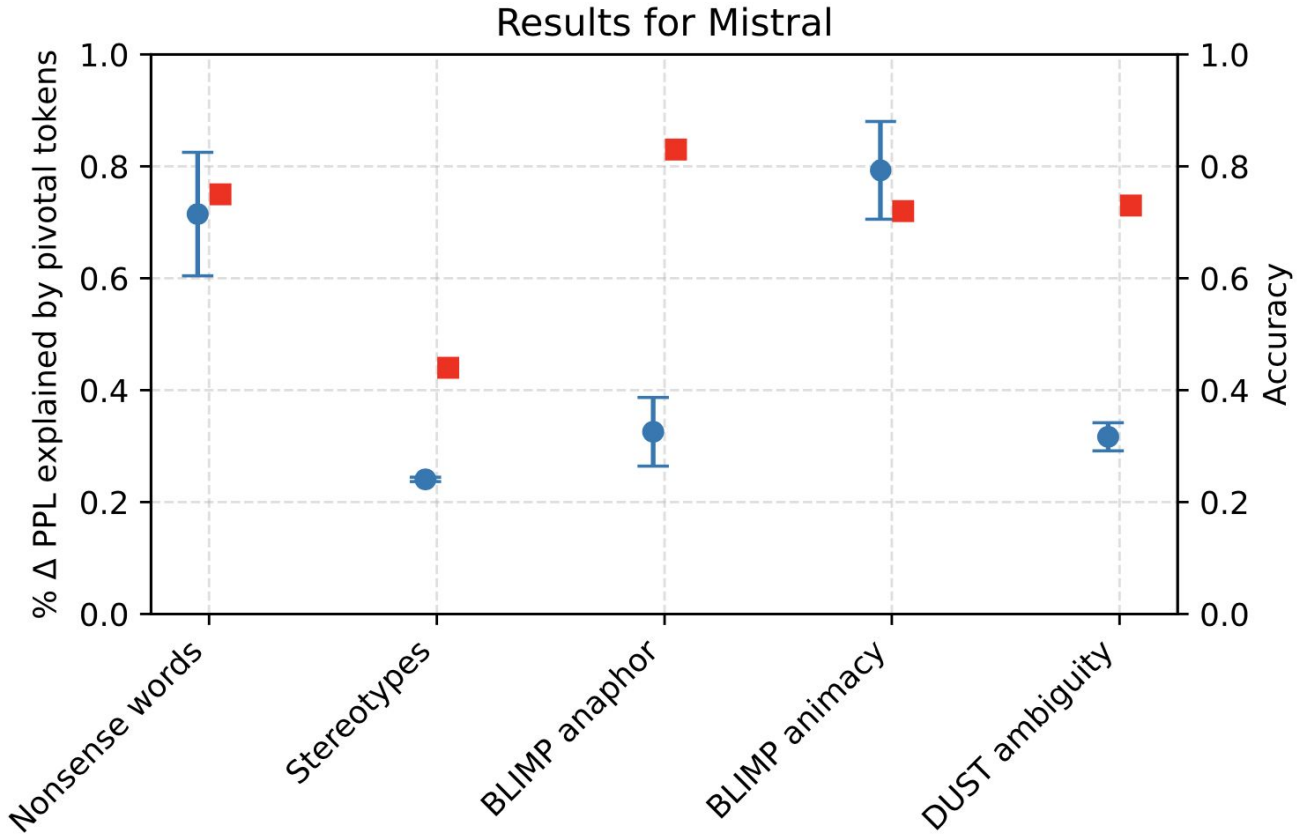
Our workflow

1. Define pivotal tokens
2. Use a minimal pair to get two prompts
3. Run the model with both prompts and record token-level perplexity
4. Take the difference in perplexity between the prompts

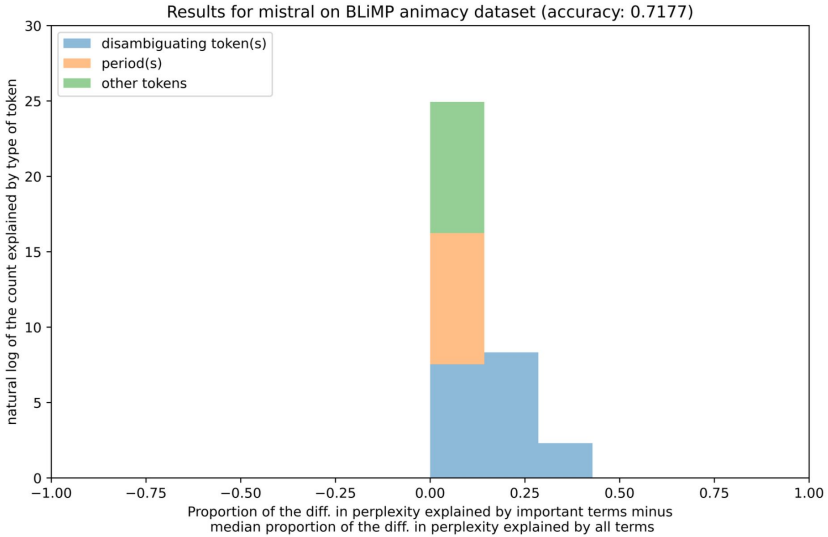
We repeated this over several different benchmarks.

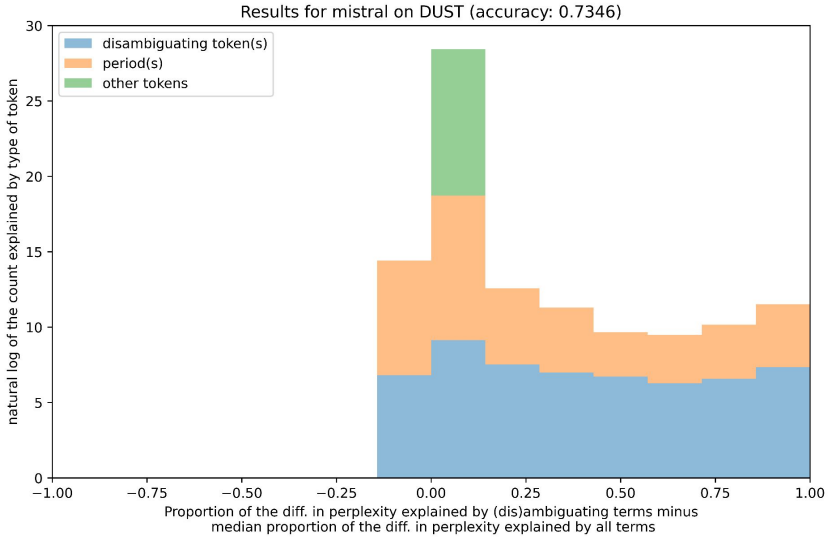
- **correct:** *"This is an ambiguous sentence: 'Andrei approached the person **with** a green chair'. This is its unambiguous counterpart: 'Andrei approached the person **who had** a green chair'."*
- **incorrect:** *"This is an ambiguous sentence: 'Andrei approached the person **who had** a green chair'. This is its unambiguous counterpart: 'Andrei approached the person **with** a green chair'."*

Results: benchmarks are too optimistic when evaluating LLMs' linguistic skills



Results: pivotal tokens are still the most influential group of tokens in these prompts





Takeaways

- confirmation bias is a problem for many XAI techniques
- we need a way to specify behavior and measure compliance
- we defined linguistic skills in terms of token-level perplexity and showed that
  - Mistral and Gemma have a lower level of linguistic skills compared to what NLP benchmarks suggest
  - Both models still seem to be influenced by the right parts of the text