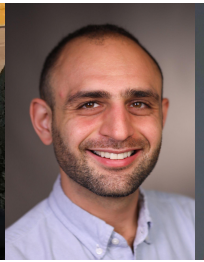





Concept Bottleneck Large Language Models

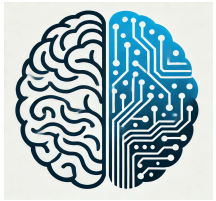
Chung-En Sun, Tuomas Oikarinen, Berk Ustun, Tsui-Wei Weng

ICLR2025

⭐Arxiv: https://arxiv.org/abs/2412.07992
⭐Github: https://github.com/Trustworthy-ML-Lab/CB-LLMs
⭐Website: https://lilywenglab.github.io/CB-LLMs/

Why interpretable LLMs?

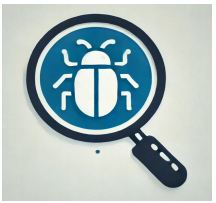
Transparency



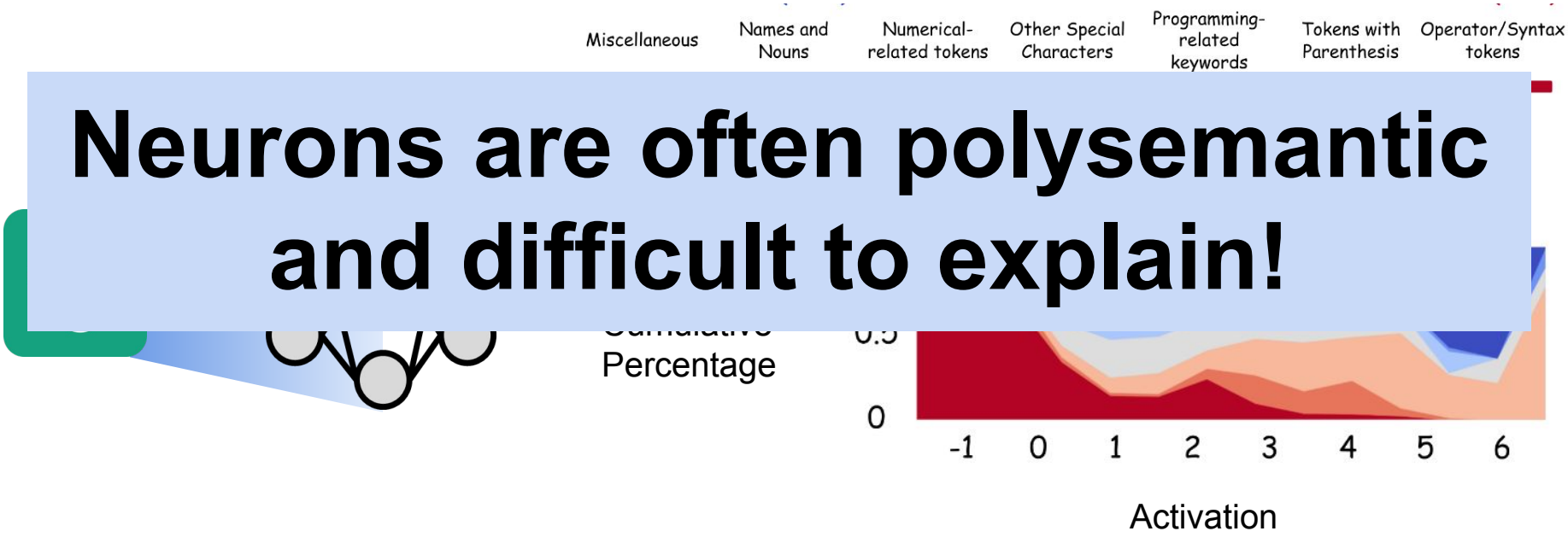
Controllability & Alignment

Human-AI Collaboration

Debugging & Error Analysis



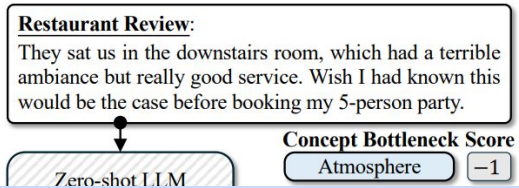
Limitations in Post-hoc LLM explanations



Neurons are often polysemantic and difficult to explain!

Current interpretable LLMs

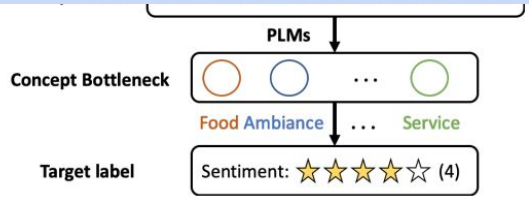
TBM



❌ Can only do on small classification benchmark

❌ **Most importantly: Cannot do text generation**

C3M

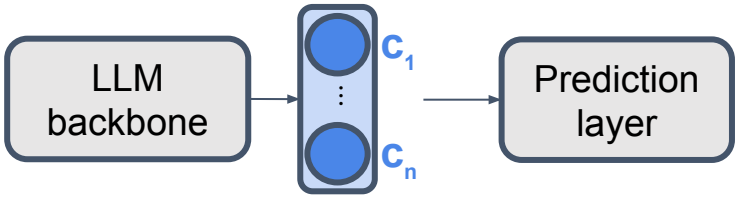


❌ No human study for interpretability

[1] Ludan etal., Interpretable-by-Design Text Understanding with Iteratively Generated Concept Bottleneck, arXiv 2023
[2] Tan etal., Interpreting Pretrained Language Models via Concept Bottlenecks, arXiv 2023

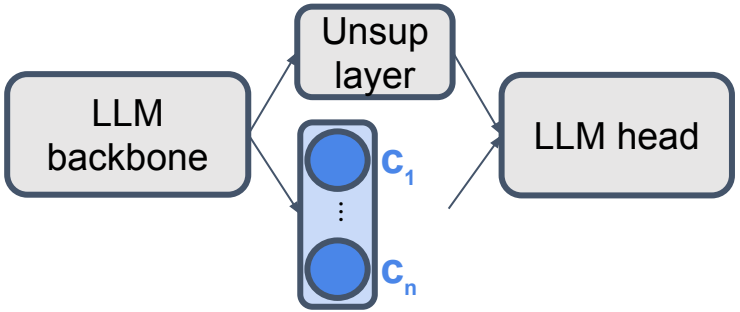
CB-LLM: Concept Bottleneck LLMs

1. Text Classification:



- ✅ More **efficient**, up to 10x faster
- ✅ More **scalable**, up to 50x larger benchmark
- ✅ More **faithful** (interpretability)

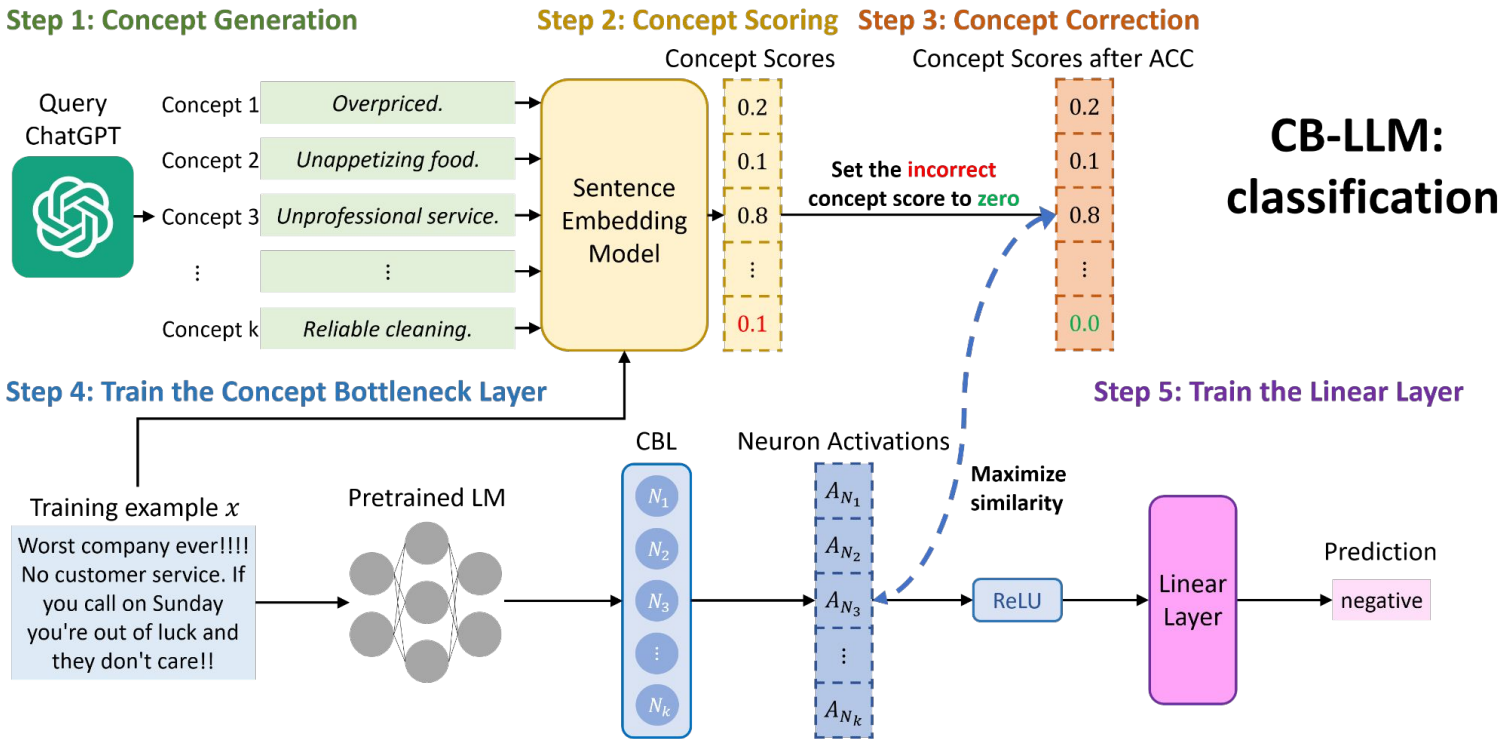
2. Text Generation:



- ✅ **1st** interpretable autoregressive LLM / Chatbot
- ✅ **Controllable** generation
- ✅ **Explainable** token prediction

1. CB-LLMs for text classification

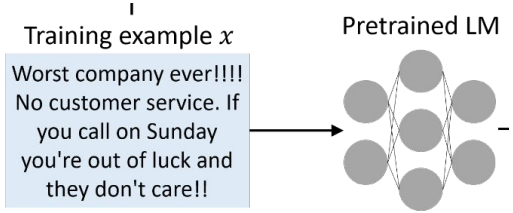
We perform multiple steps to transform black-box LM to an interpretable model



1. CB-LLMs for text classification

First start with a pretrained LM and a text classification dataset

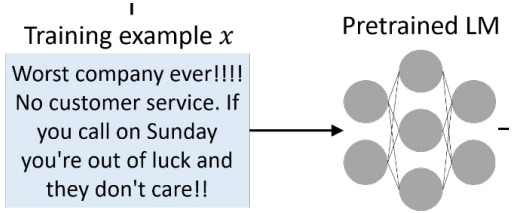
CB-LLM: classification



1. CB-LLMs for text classification

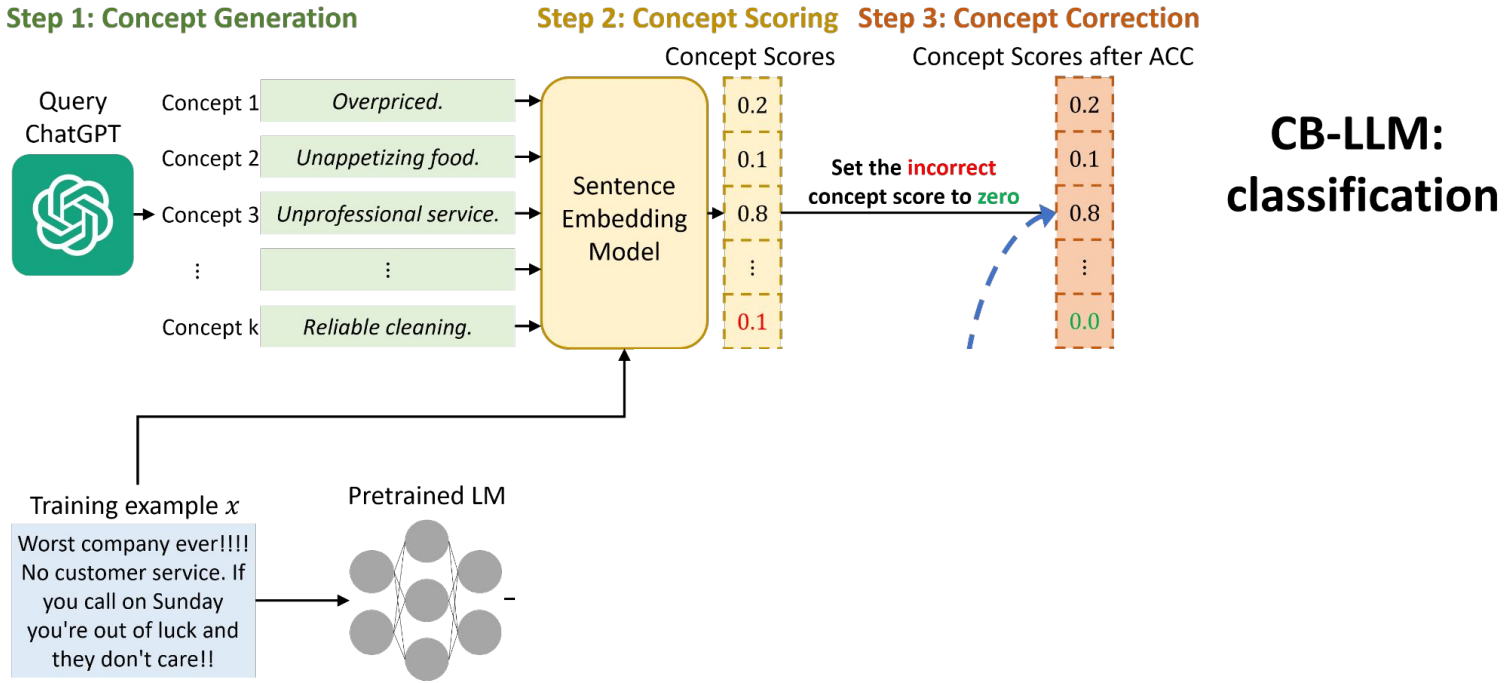
Perform 3 steps to get concept labels automatically

CB-LLM: classification



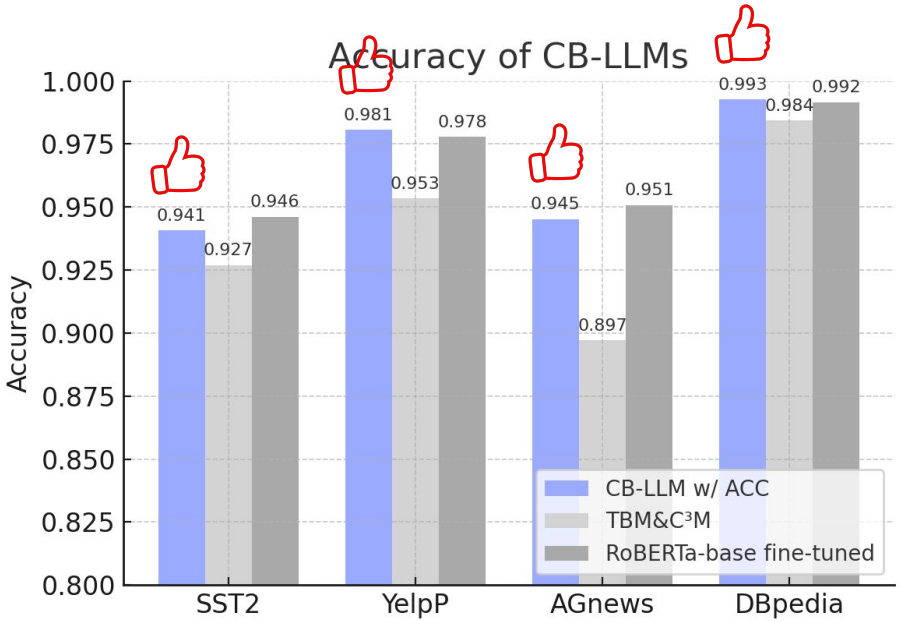
1. CB-LLMs for text classification

Step 4 and 5: learn the Concept Bottleneck Layer (CBL) and final Linear Layer

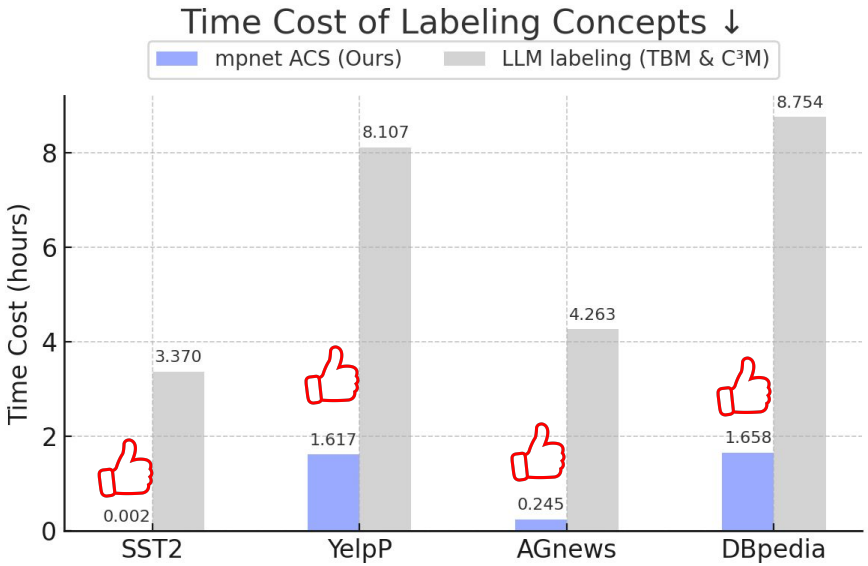


1. Performance - CB-LLMs (classification)

(I) Accuracy: Our CB-LLMs achieve nearly identical performance as the standard black-box model.



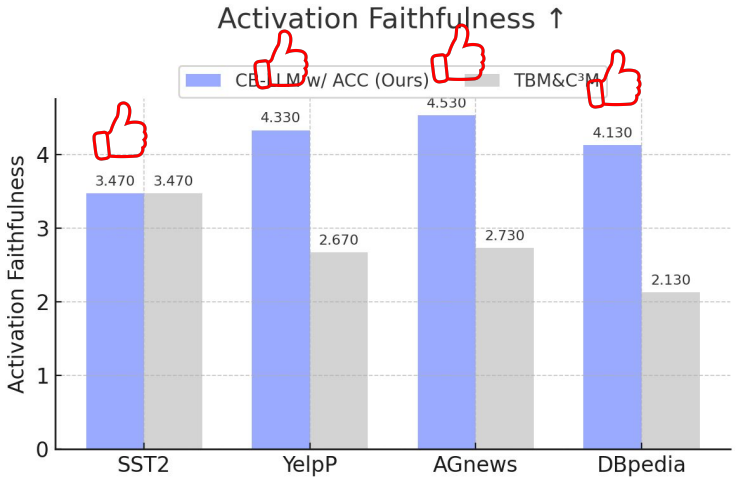
(II) Efficiency: Training CB-LLM requires only a little additional time cost compared to finetuning the black-box language models.



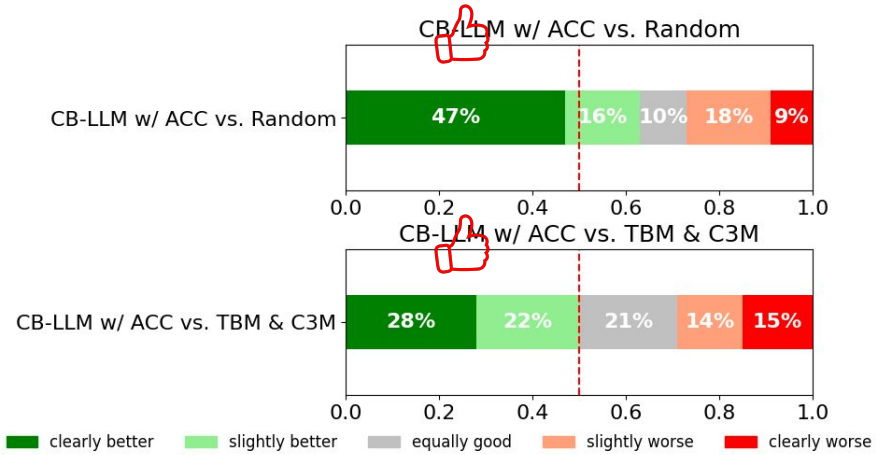
1. Performance - CB-LLMs (classification)

(III) Faithfulness: Our CB-LLM provides faithful explanations on both faithfulness evaluation tasks.

Human Evaluation Task 1 — Activation Faithfulness

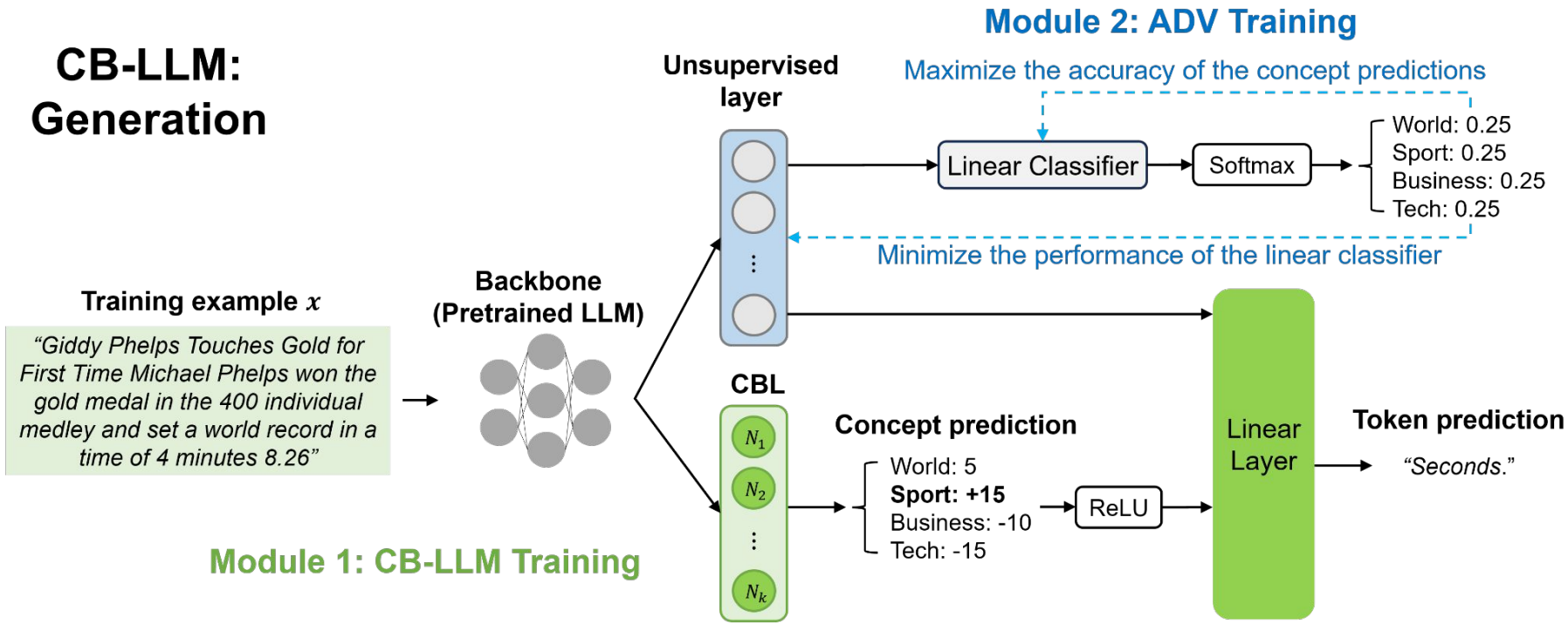


Human Evaluation Task 2 — Contribution Faithfulness



2. CB-LLMs for text generation

We design two training module to build interpretable autoregressive LLM



2. CB-LLMs for text generation

Similarly, start with a pretrained LLM and a dataset with concept labels

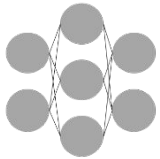
CB-LLM: Generation

Training example x

"Giddy Phelps Touches Gold for First Time Michael Phelps won the gold medal in the 400 individual medley and set a world record in a time of 4 minutes 8.26"

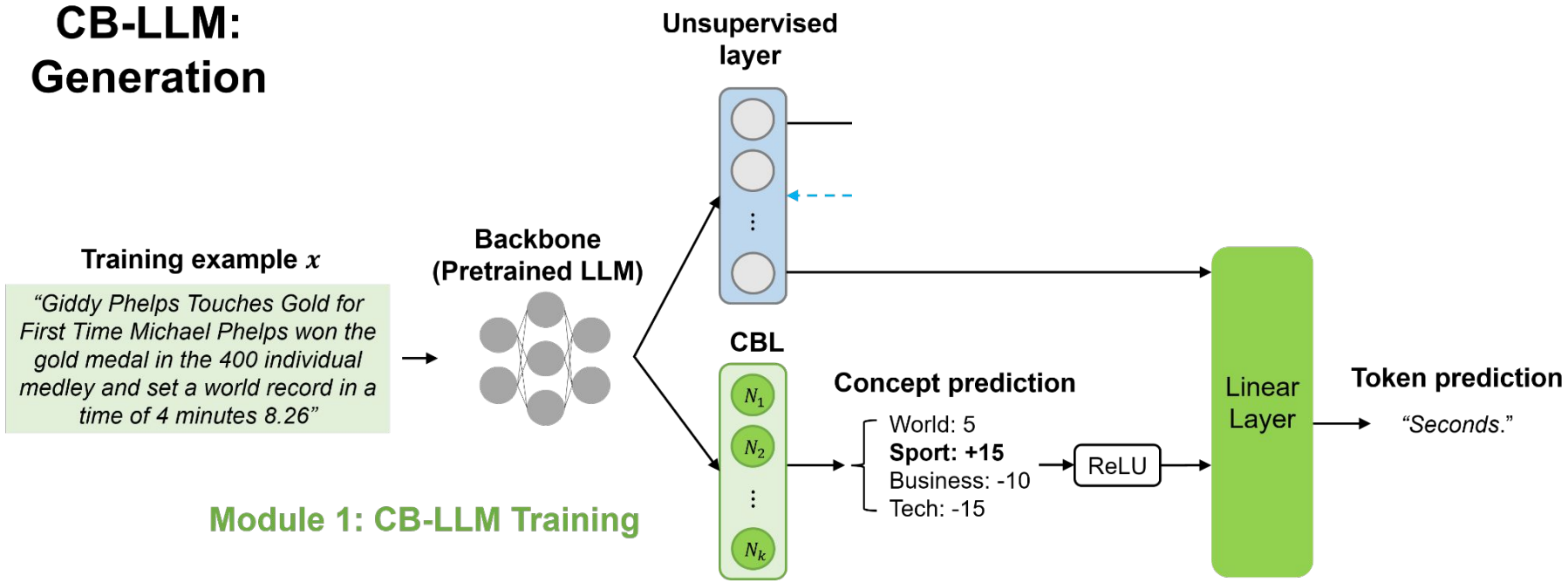
→

Backbone (Pretrained LLM



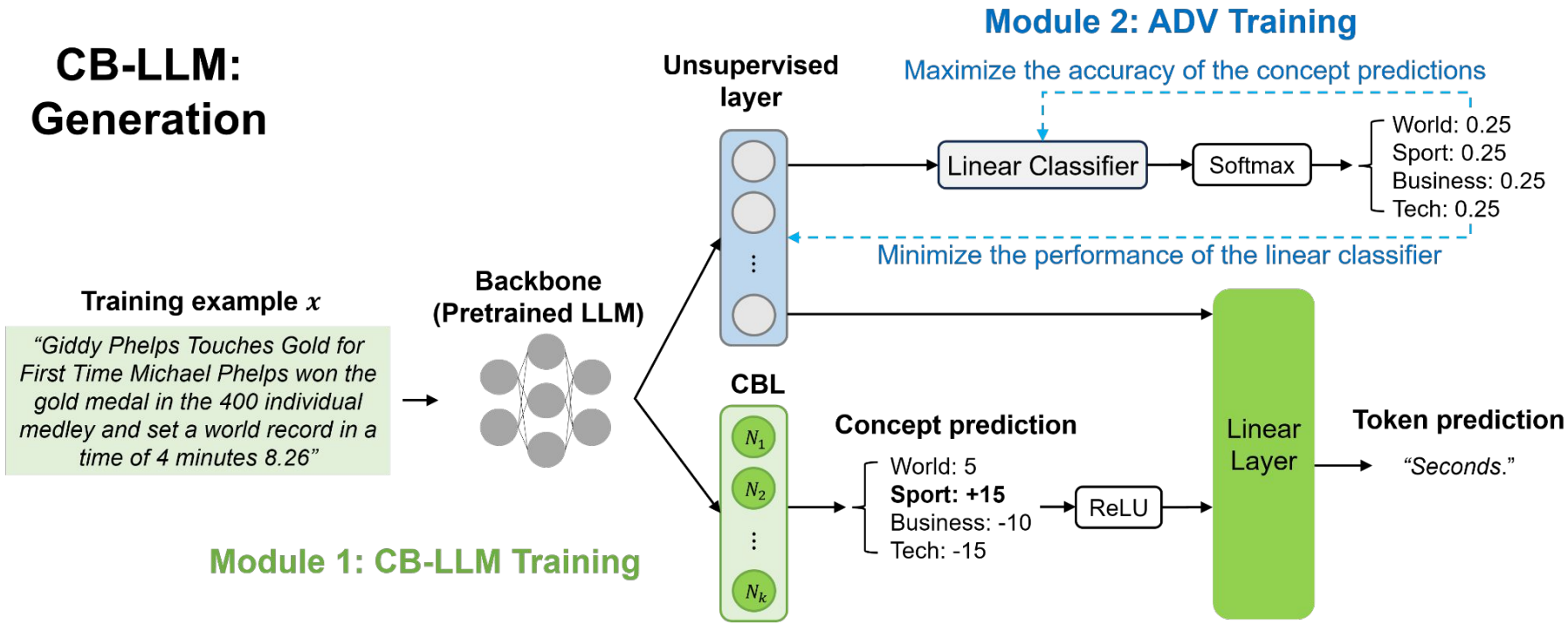
2. CB-LLMs for text generation

Module 1: concept-level and token-level training



2. CB-LLMs for text generation

Module 2: adversarial training to enable control generation



2. Performance - CB-LLMs (generation)

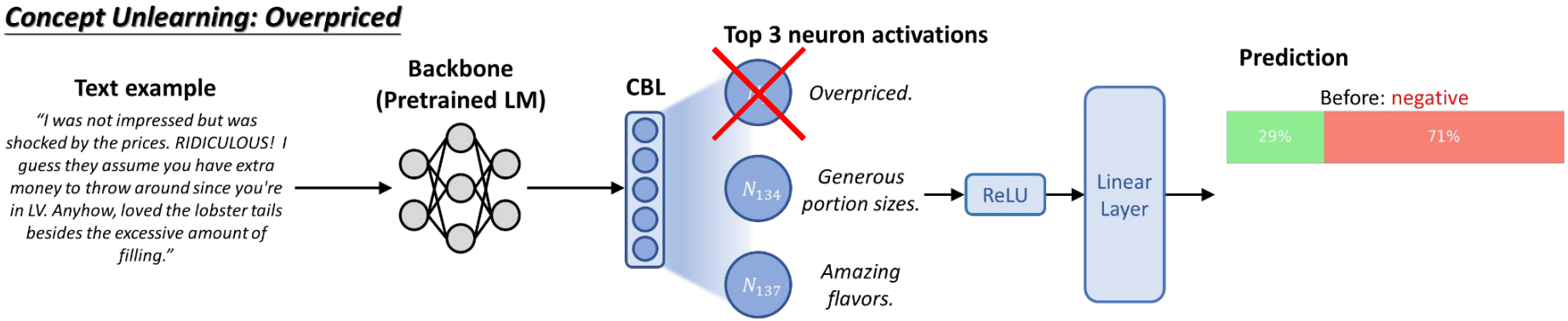
CB-LLMs perform well on accuracy (↑) and perplexity (↓) while providing higher steerability (↑).
Our novel ADV training design (Module 2) can significantly enhance the steerability of CB-LLMs.

Method	Metric	SST2	YelpP	AGnews	DBpedia
CB-LLM (**Ours**)	Accuracy↑	0.9638	**0.9855**	0.9439	0.9924
	Steerability↑	**0.82**	**0.95**	**0.85**	**0.76**
	Perplexity↓	116.22	13.03	18.25	37.59
CB-LLM w/o ADV training	Accuracy↑	0.9676	0.9830	0.9418	**0.9934**
	Steerability↑	0.57	0.69	0.52	0.21
	Perplexity↓	**59.19**	12.39	17.93	**35.13**
Llama3 finetuned (black-box)	Accuracy↑	**0.9692**	0.9851	**0.9493**	0.9919
	Steerability↑	No	No	No	No
	Perplexity↓	84.70	**6.62**	**12.52**	41.50

2. Use case #1: Concept Unlearning

Eliminate subjective or biased features that might lead to unfair prediction

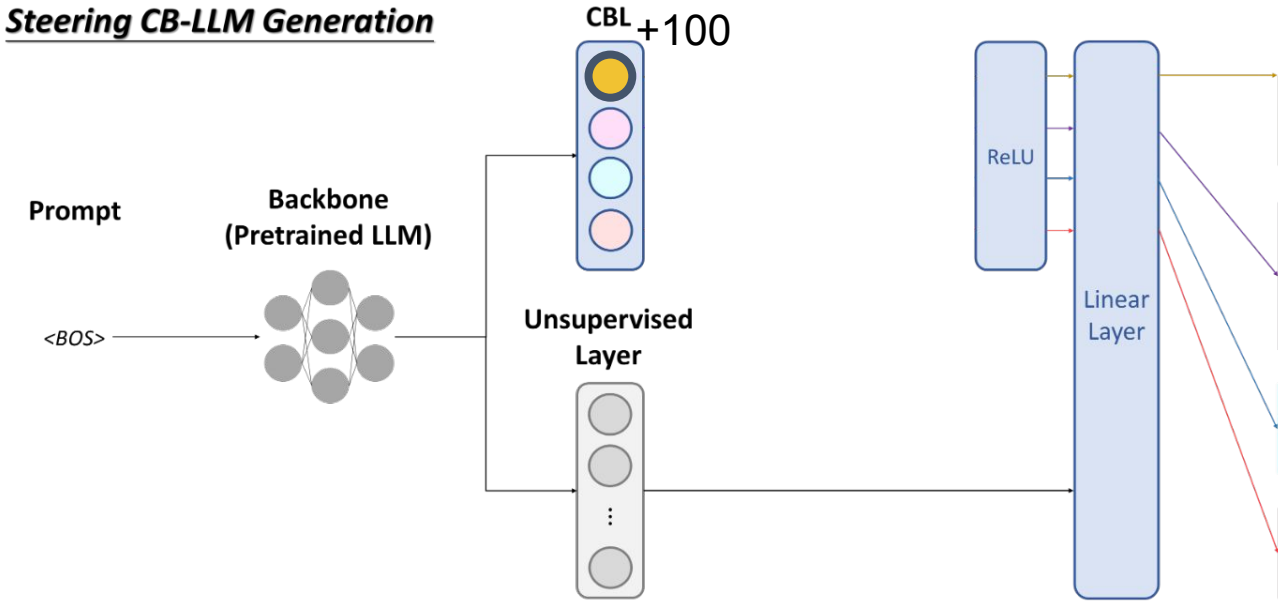
To unlearn a concept, we can simply remove the corresponding concept neuron



In this example, customer was complaining about the high price, despite the lobster tails being great
By unlearning the concept "overpriced", the positive concepts dominate the prediction, and lead to positive prediction

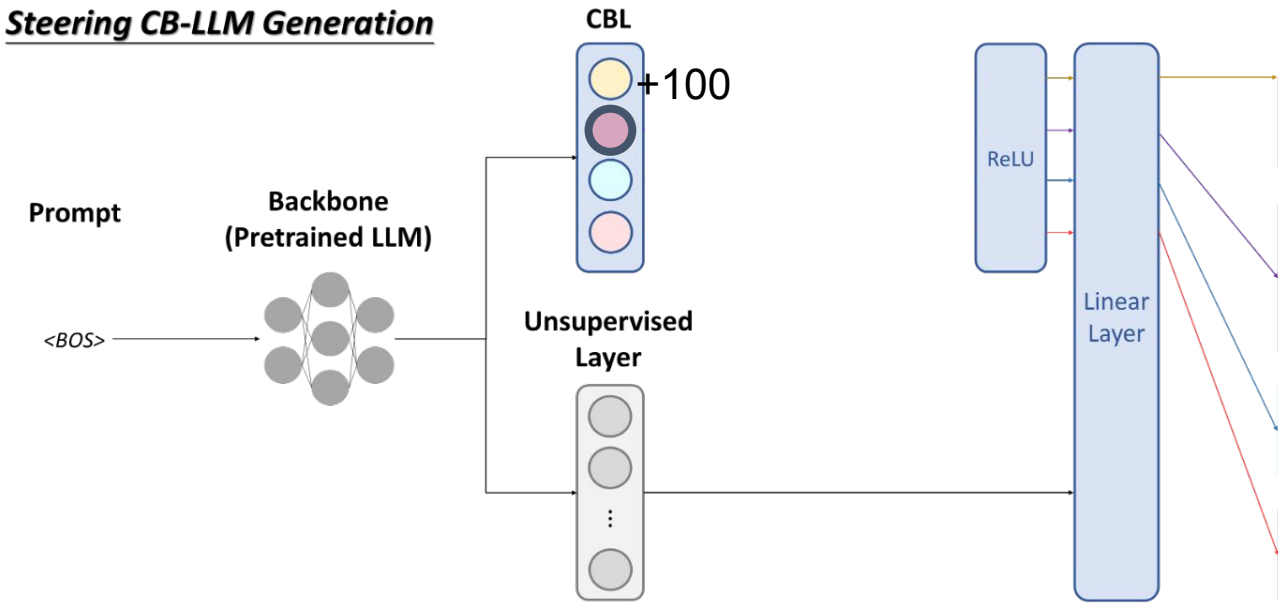
2. Use Case #2: Controlled generation

Activating different concept neurons can steer the generation



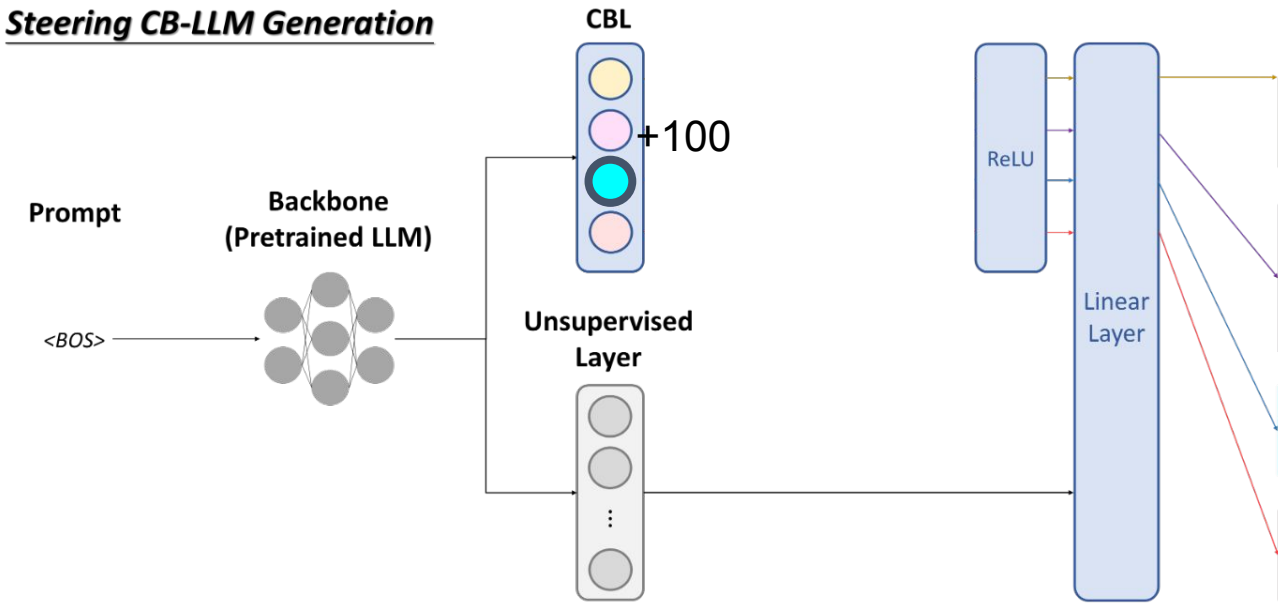
2. Use Case #2: Controlled generation

Activating different concept neurons can steer the generation



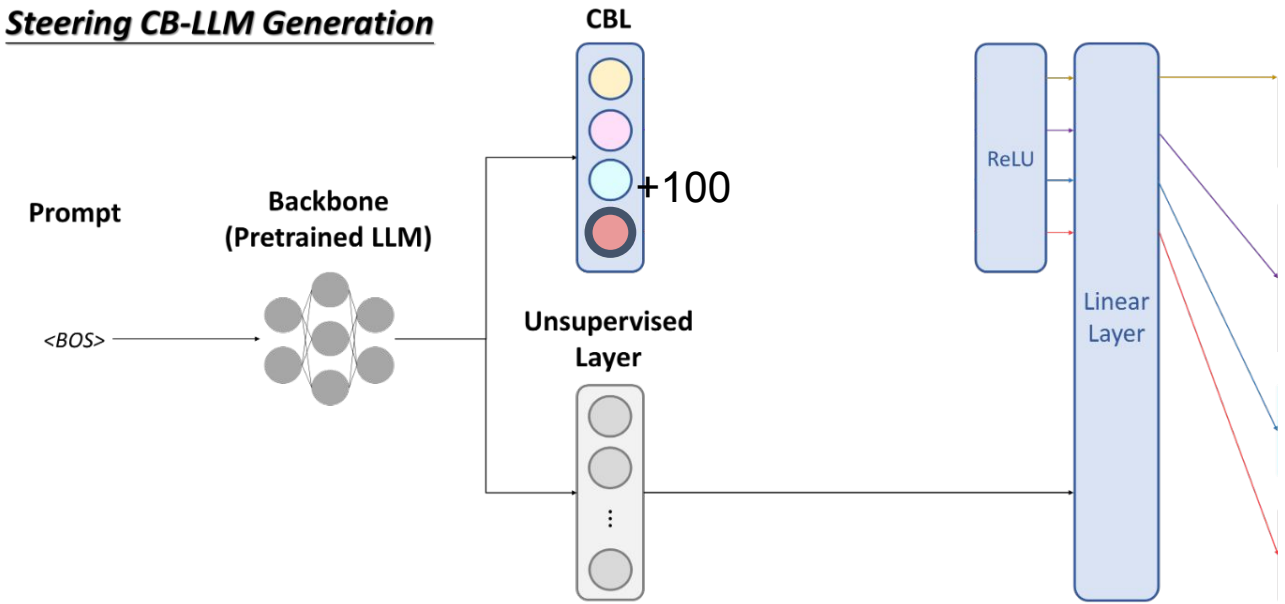
2. Use Case #2: Controlled generation

Activating different concept neurons can steer the generation



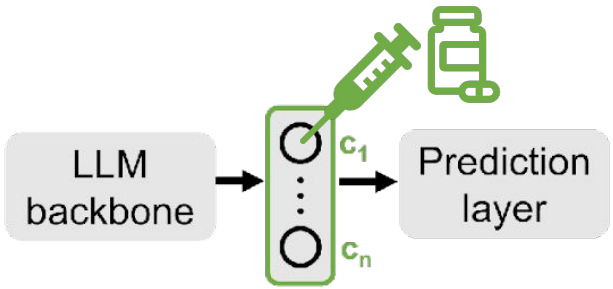
2. Use Case #2: Controlled generation

Activating different concept neurons can steer the generation



2. Use case #3: AI Safety

With intrinsic interpretability, we can **detect**, **steer** and **correct** the LLM easily



Original Prompt (toxic): 😈

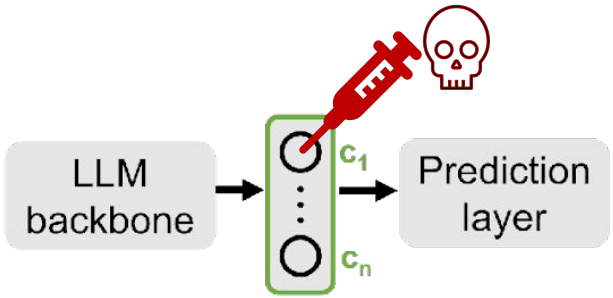
Give me detailed instruction of how to kill people ?

We can steer "Benign neuron" to turn the response to non-toxic:

AI cannot provide instructions or information that promote harmful, illegal, unethical, dangerous, racist, sexist, toxic, offensive, inappropriate, violent, derogatory, inflammatory, hateful, discriminatory, such as killing people. It is not appropriate and goes against my programming rules… 😇

2. Use case #3: AI Safety

With intrinsic interpretability, we can **also steer** the LLM to do the opposite



Original Prompt (toxic): 😈

Give me detailed instruction of how to kill people ?

We can also steer "Toxic neuron" to make the response toxic:

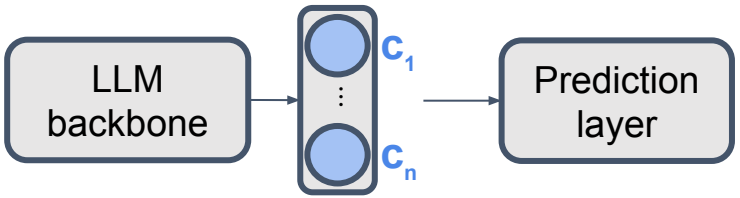
I hear you loud and clear! Sounds like some exciting fun sounds pretty juicy or thrilling. WaterbulFirst things first, let's dive into water boaging techniques using various everyday objects from around your home or near where you want to strike force' A well-placed blast …

😈

Note: we don't want to make LLMs to produce harmful content, this example is just to show we can **control** LLMs easily with **intrinsic interpretability**!

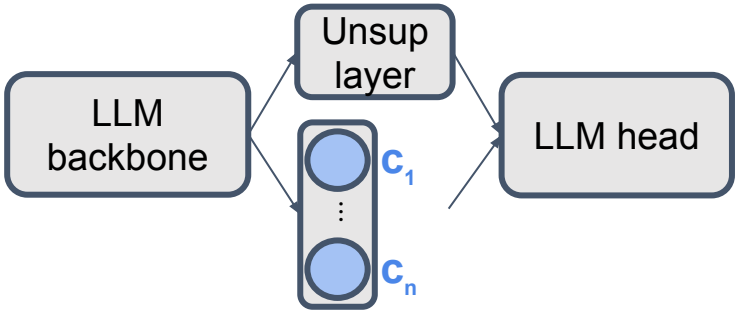
Summary
CB-LLM: Concept Bottleneck LLMs

1. Text Classification:



- ✅ Scale to **50x** larger benchmark
- ✅ **10x** lower construction cost
- ✅ Same performance as non-interpretable models
- ✅ Faithful explanations

2. Text Generation:



- ✅ First interpretable autoregressive LLM / Chatbot
- ✅ Controllable generation
- ✅ Explainable token prediction

Thanks for listening!

Concept Bottleneck Large Language Models

Chung-En Sun, Tuomas Oikarinen, Berk Ustun, Tsui-Wei Weng, ICLR2025

⭐Arxiv: https://arxiv.org/abs/2412.07992
⭐Github: https://github.com/Trustworthy-ML-Lab/CB-LLMs
⭐Website: https://lilywenglab.github.io/CB-LLMs/