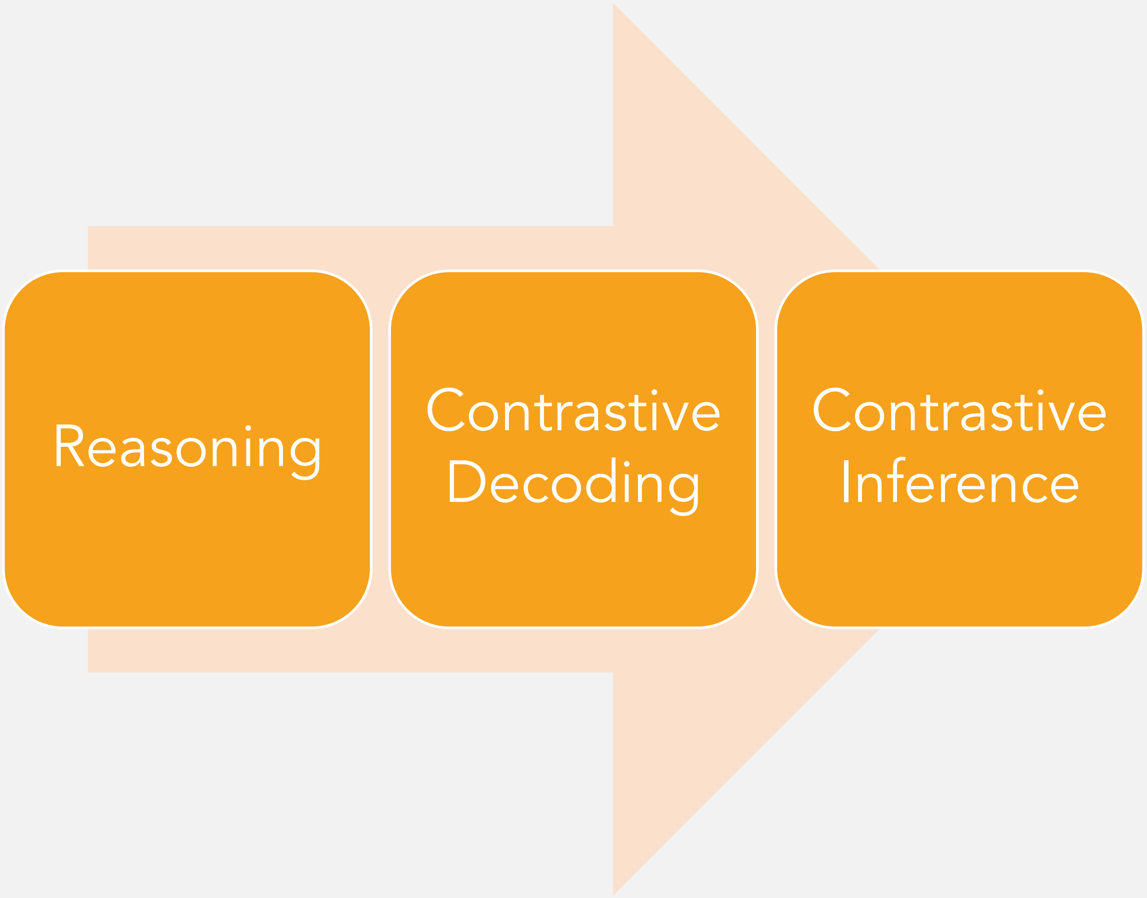
CONTRASTIVE INFERENCE

Sean O'Brien

seobrien@ucsd.edu

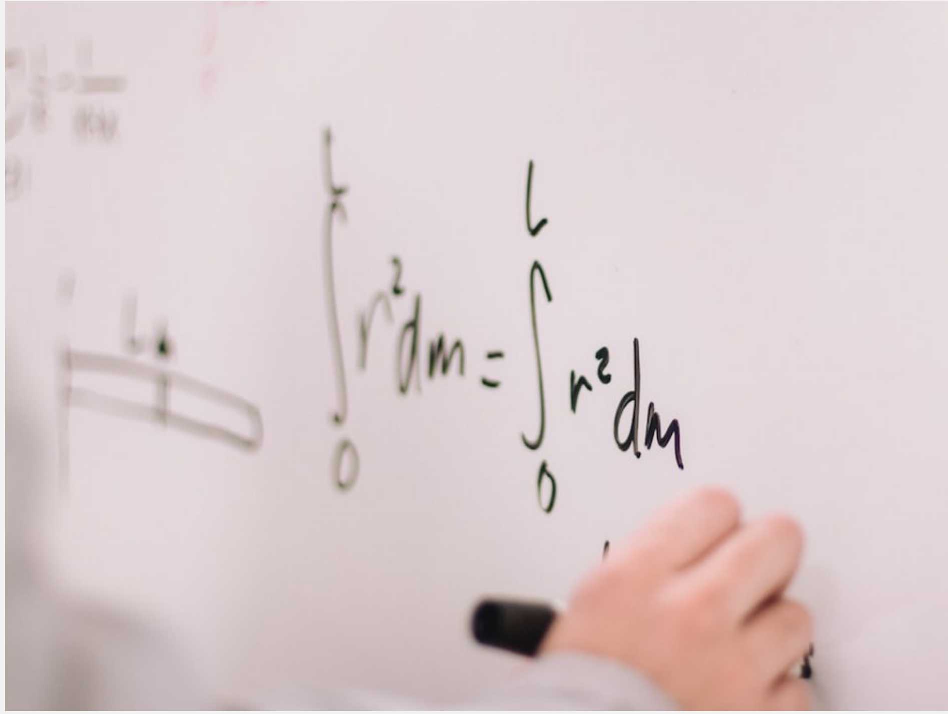
TALK STRUCTURE



REASONING WITH LLMS

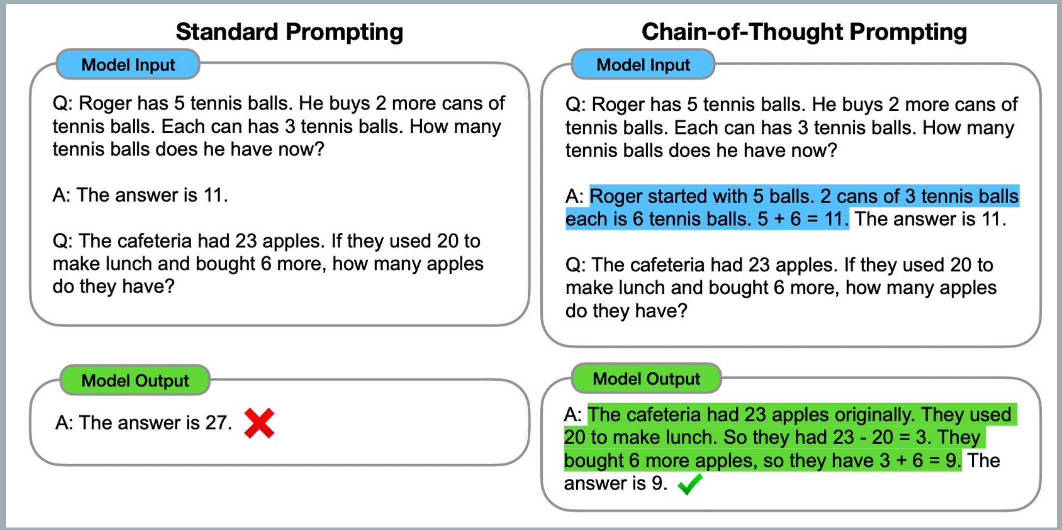
MOTIVATION

✅



❌

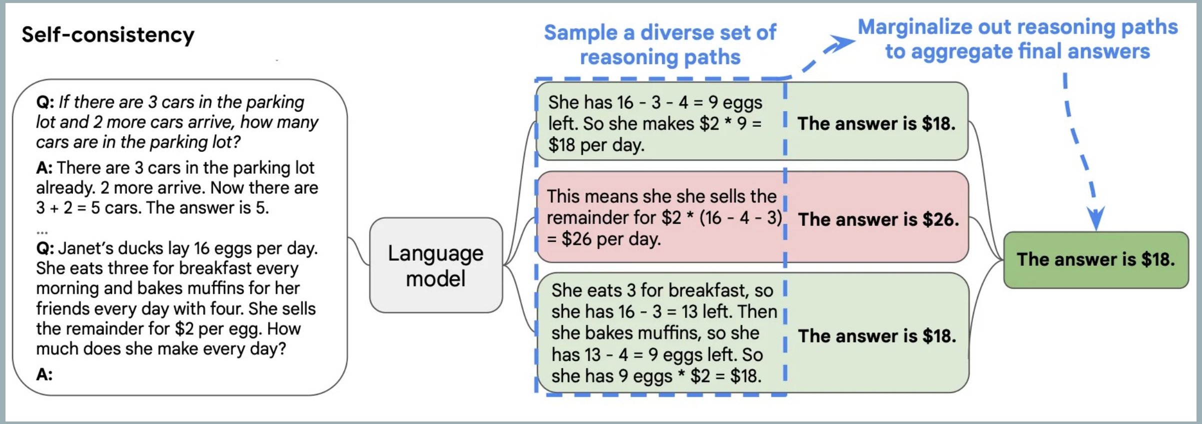
CHAIN-OF-THOUGHT[1]



- **Motivation:** Decomposing tasks into intermediate steps makes them easier.
- **Idea:** Prompt a model to output a full reasoning chain before its answer.
- **Result:** Performance soars almost universally, given models are large enough.
- **Takeaway:** Large models can exhibit stronger reasoning capacities based on how we prompt them.
  - More abstractly, the **decoding method** / **prompt** is an important limiting factor for performance.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

SELF-CONSISTENCY[1]



- **Motivation:** Multiple reasoning paths could take you to the right answer.
- **Idea:** Sample multiple full generations from a model, then aggregate the final answers.
- **Result:** The best method is to just take a simple majority vote from the answers.
  - Results improve drastically and reliably.
- **Takeaway:** Sampling can be useful for reasoning, but only in conjunction with SC.
  - Parallelizable, but takes a lot of extra compute.
  - Hard for solving open-ended questions with answers that are difficult to group together.

1. Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language

DECODING METHODS – THE SPLIT

REASONING

- **Greedy decoding preferred**
- Most work done on the prompting and augmentation level
  - Chain-of-thought prompting
  - Program-aided language models
  - LLM prompt optimizer

OPEN-ENDED GENERATION

- **Sampling methods preferred**
- Truncated sampling schemes work best
  - Top-*k* sampling
  - Nucleus sampling
  - Typical sampling

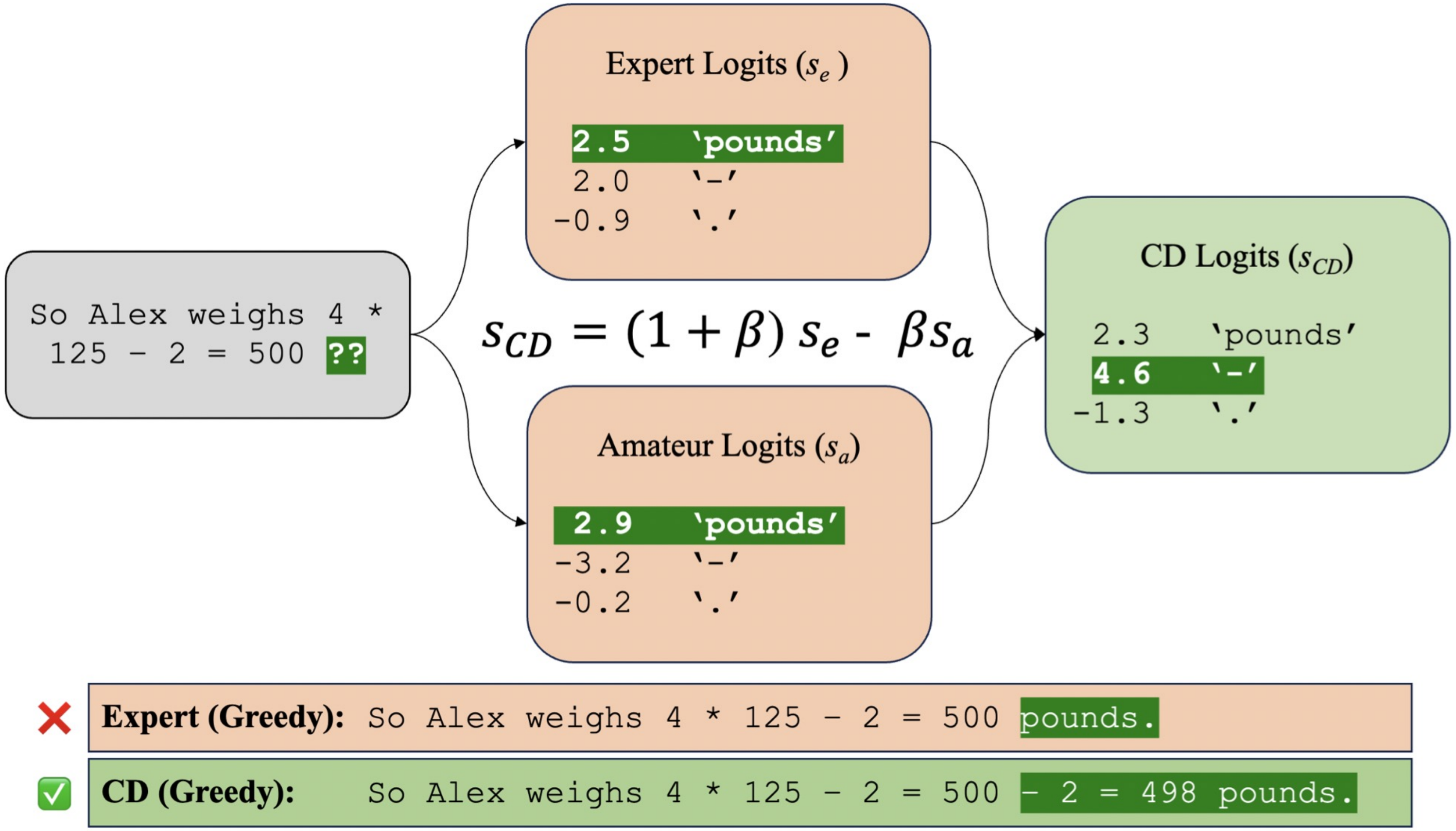
General-Purpose Method?

CONTRASTIVE DECODING

INDUCTIVE BIAS

Large language models are better than small language models.

CD: VISUALIZED



Sean O'Brien and Mike Lewis. Contrastive Decoding Improves Reasoning in Large Language Models, 2022.

CD OBJECTIVE

$$\text{CD-score}(x_i; x_{<i}) \tag{3}$$
$$= \begin{cases} \boxed{\log \frac{p_{\text{EXP}}(x_i|x_{<i})}{p_{\text{AMA}}(x_i|x_{<i})}}, & \text{if } x_i \in \mathcal{V}_{\text{head}}(x_{<i}), \\ -\inf, & \text{otherwise.} \end{cases}$$

- CD replaces the standard decoding objective

$$\max_w p_{EXP}(w)$$

with

$$\max_w p_{EXP}(w)/p_{AMA}(w)$$

- The original paper greedily optimizes this.

Challenges

- Instability associated with tokens the amateur considers highly unlikely
- Breaks down when the amateur and expert agree

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text Generation as Optimization, 2022.

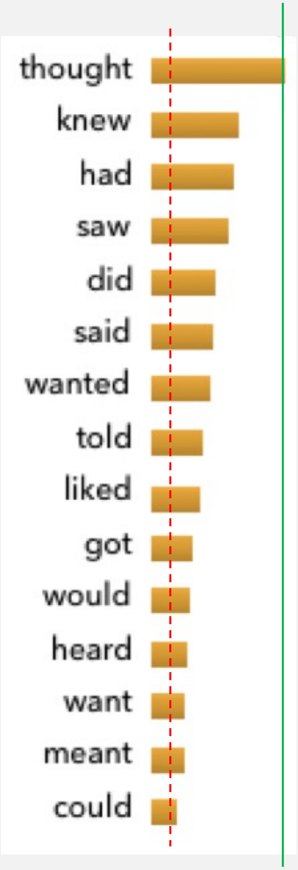
α-MASKING

$$\mathcal{V}_{\text{head}}(x_{<i}) = \tag{1}$$
$$\{x_i \in \mathcal{V} : p_{\text{EXP}}(x_i \mid x_{<i}) \geq \alpha \max_w p_{\text{EXP}}(w|x_{<i})\}$$

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text Generation as Optimization, 2022.

- We want to restrict candidate tokens based on what the expert finds reasonably likely
- Other truncation techniques can break down:
  - Top-*k* masking can include low-probability tokens
  - Nucleus sampling can eliminate viable candidates in high-entropy situations
- α -masking is another adaptive masking strategy
- Fairly insensitive parameter, but 0.1 tends to work.

α-MASKING



Holtzman et al 2020. "The Curious Case of Neural Text Degeneration."

MODIFIED METHOD

1. Determine α-mask.

$$V_{valid} = \{j \in V, s_e^{(j)} \geq \log \alpha + \max_{k \in V} s_e^{(k)}\}$$

2. Subtract amateur logits.

$$s_{CD}^{(i)} = \begin{cases} (1+\beta)s_e^{(i)} - \beta s_a^{(i)} & i \in V_{valid} \\ -\infty & i \notin V_{valid} \end{cases}$$

Sean O'Brien and Mike Lewis. Contrastive Decoding Improves Reasoning in Large Language Models, 2022.

- The pre-contrast amateur and expert temperatures are slightly unintuitive.
- We keep α the same, but simplify the mask calculation.
- We introduce β, which is the strength of the contrastive penalty.
  - To keep it orthogonal with sampling temperature, we scale the expert logits up by $(1+\beta)$
- Results are sensitive to β
  - 0.5 works well for most tasks, but it depends on the gap between the expert and amateur

PYTORCH IMPLEMENTATION

Algorithm 2: Our formulation

```
# expert_logits - unnormalized scores from the expert model
# amateur_logits - unnormalized scores from the amateur model
# alpha - masking threshold
# beta - expert-amateur tradeoff parameter

cutoff = log(alpha) + expert_logits.max(dim=-1, keepdim=True).values
diffs = (1 + beta)*expert_logits - beta*amateur_logits
cd_logits = diffs.masked_fill(expert_logits < cutoff, -float('inf'))
```

Sean O'Brien and Mike Lewis. Contrastive Decoding Improves Reasoning in Large Language Models, 2022.

RESULTS

CD (ORIGINAL)

Humans prefer generations from CD to sampling methods

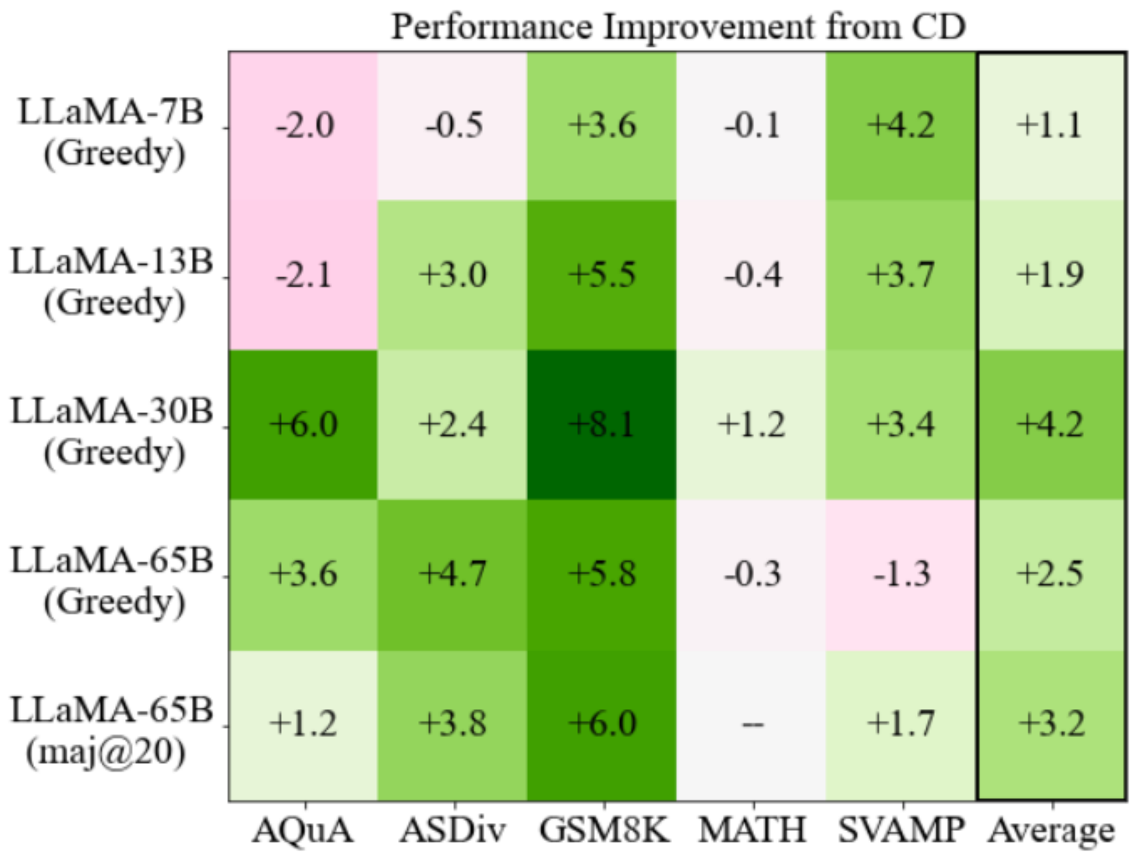
CD tends to improve diversity and coherence

Results are best when there is a large expert/amateur gap

	CD	Baseline	coherence: CD is better	coherence: same	coherence: Baseline is better	fluency: CD is better	fluency: same	fluency: Baseline is better
wikitext	CD (GPT-2 XL)	nucleus (GPT-2 XL)	**0.714***	0.083	0.202	**0.548**	0.083	0.369
	CD (GPT-2 XL)	typical (GPT-2 XL)	**0.887***	0.046	0.067	**0.703***	0.082	0.215
	CD (OPT-13B)	nucleus (OPT-13B)	**0.556**	0.202	0.242	**0.419**	0.197	0.384
	CD (OPT-13B)	typical (OPT-13B)	**0.773***	0.106	0.121	**0.687***	0.152	0.162
wikinews	CD (GPT-2 XL)	nucleus (GPT-2 XL)	**0.708***	0.042	0.25	**0.583***	0.12	0.297
	CD (GPT-2 XL)	typical (GPT-2 XL)	**0.771***	0.151	0.078	**0.755***	0.151	0.094
	CD (OPT-13B)	nucleus (OPT-13B)	**0.585***	0.221	0.195	**0.518**	0.123	0.359
	CD (OPT-13B)	typical (OPT-13B)	**0.693***	0.099	0.208	**0.49**	0.297	0.214
story	CD (GPT-2 XL)	nucleus (GPT-2 XL)	**0.636***	0.045	0.318	0.404	0.106	**0.49**
	CD (GPT-2 XL)	typical (GPT-2 XL)	**0.506**	0.256	0.238	**0.387**	0.363	0.25
	CD (OPT-13B)	nucleus (OPT-13B)	**0.616***	0.101	0.283	**0.449**	0.293	0.258
	CD (OPT-13B)	typical (OPT-13B)	**0.626***	0.202	0.172	**0.52**	0.212	0.268

CD

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text Generation as Optimization, 2022.



CD (MATH)

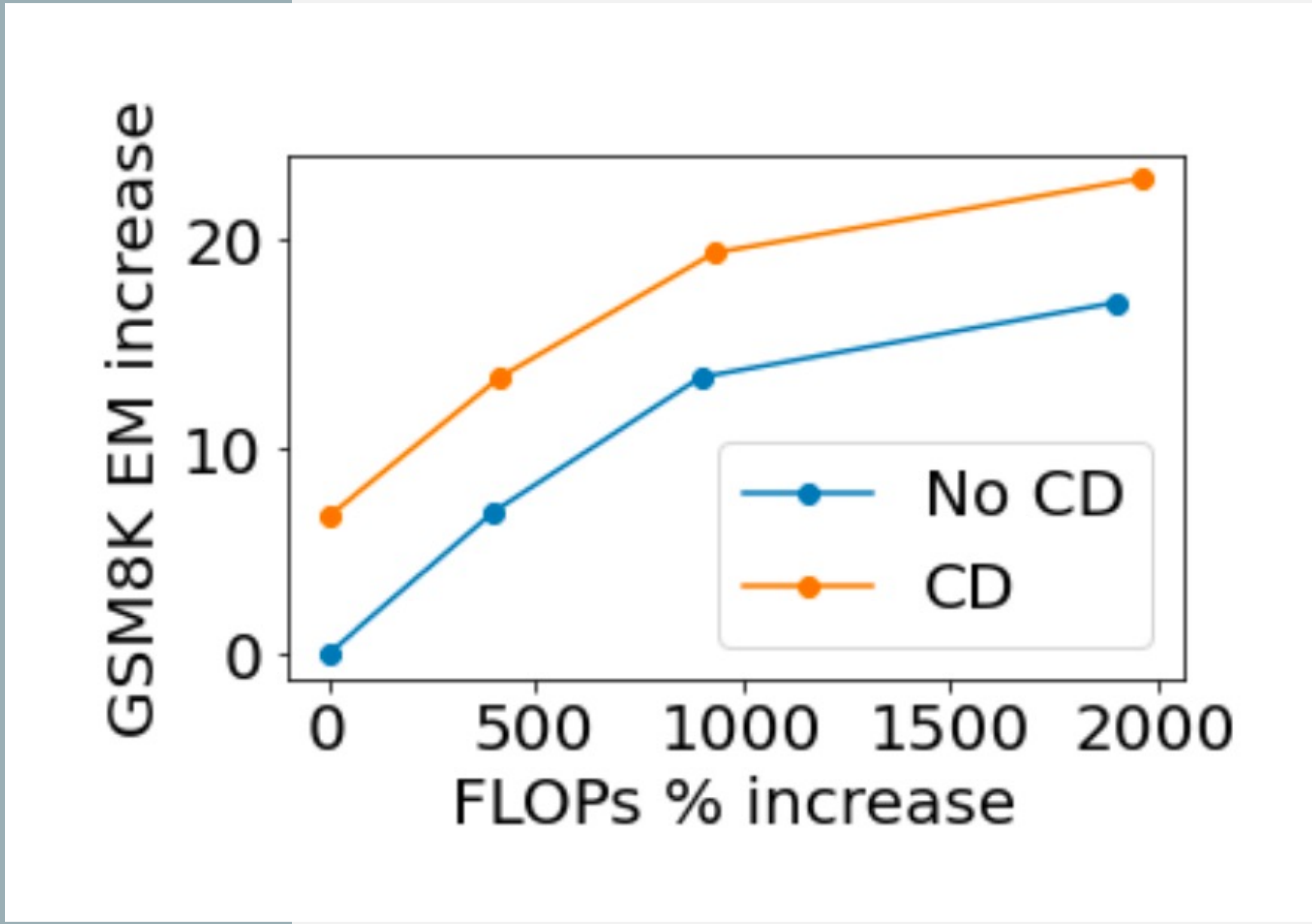
- Performance tends to improve on math tasks
- Doesn’t help on problems that the expert can’t solve either
  - AQuA for 7B and 13B models
  - MATH for all models
- Combines well with self-consistency

Sean O’Brien and Mike Lewis. Contrastive Decoding Improves Reasoning in Large Language Models, 2022.

CD + SC

CD combines with self-consistency to be very strong

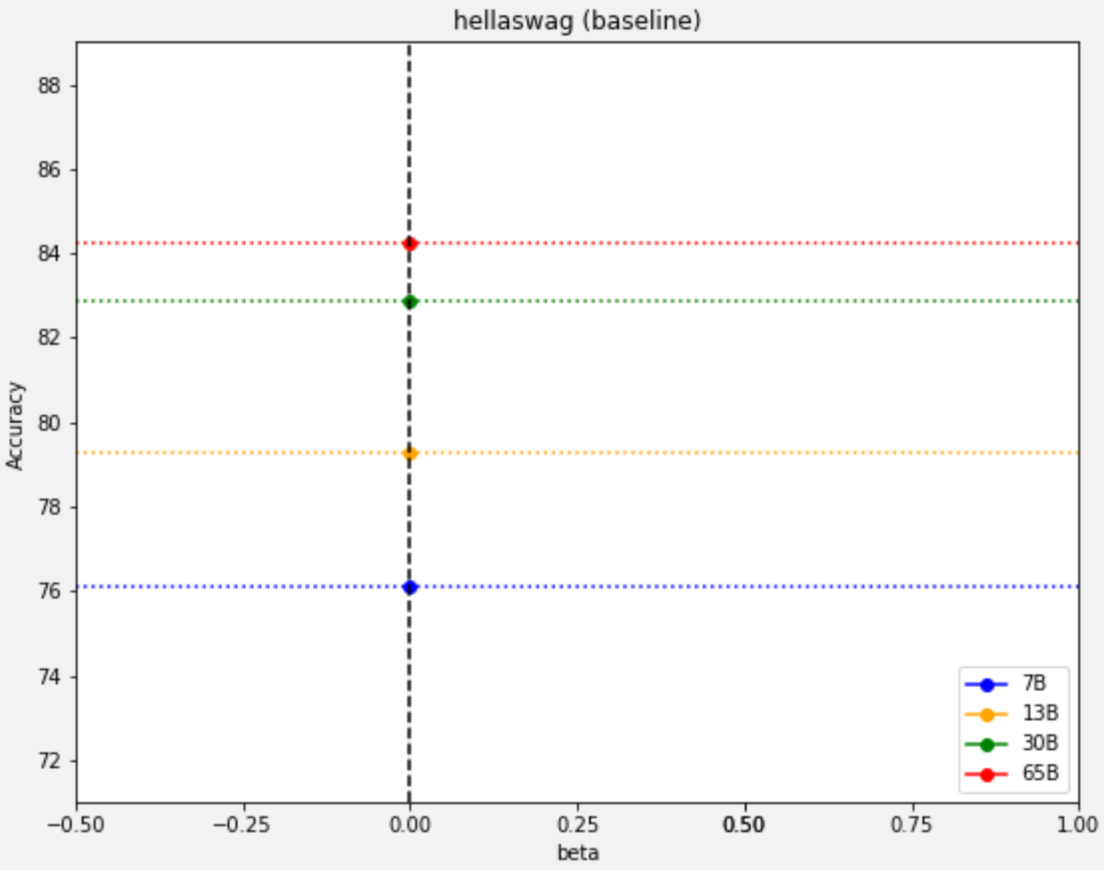
CD provides a much more compute-efficient benefit than self-consistency



Sean O'Brien and Mike Lewis. Contrastive Decoding Improves Reasoning in Large Language Models, 2022.

HELLASWAG

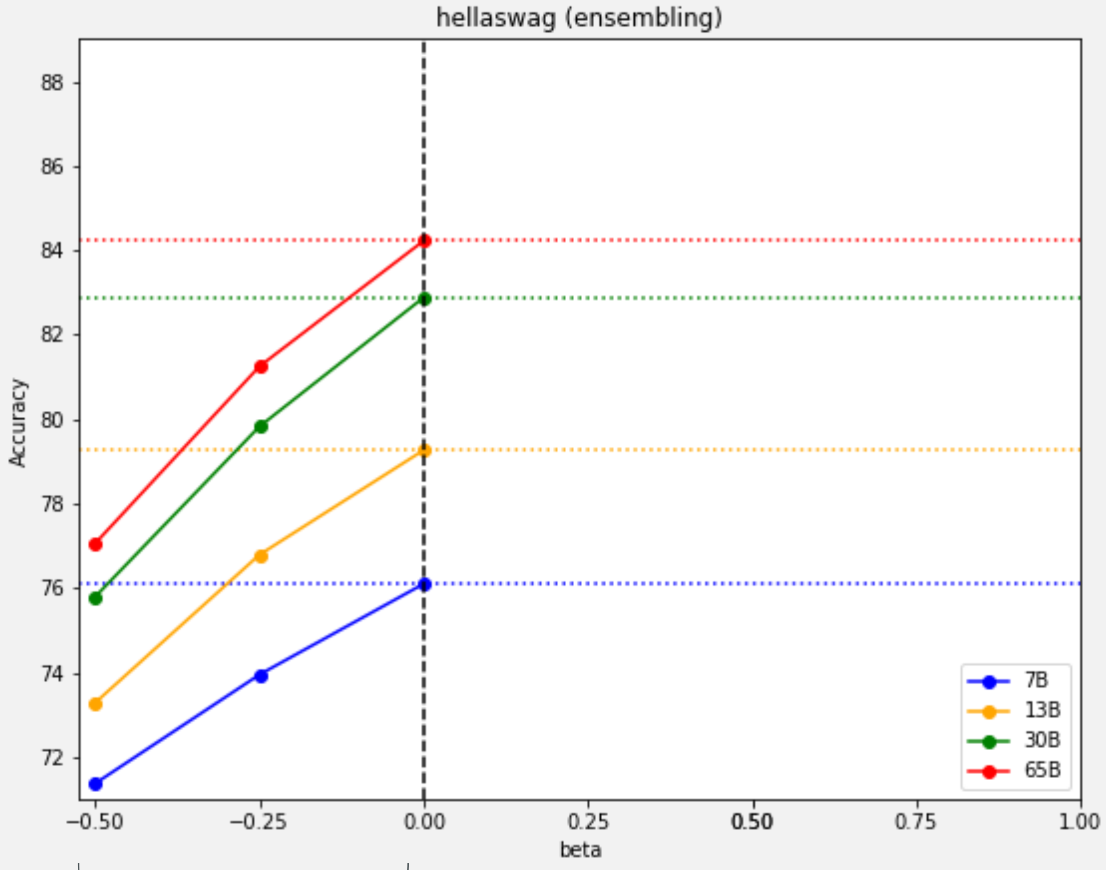
Model	Score
LLaMA **65B**	84.2
LLaMA **2**	85.3
ChatGPT	85.5
PaLM **2**	86.8



Sean O'Brien and Mike Lewis. Contrastive Decoding Improves Reasoning in Large Language Models, 2022.

HELLASWAG

Model	Score
LLaMA **65B**	84.2
LLaMA **2**	85.3
ChatGPT	85.5
PaLM **2**	86.8

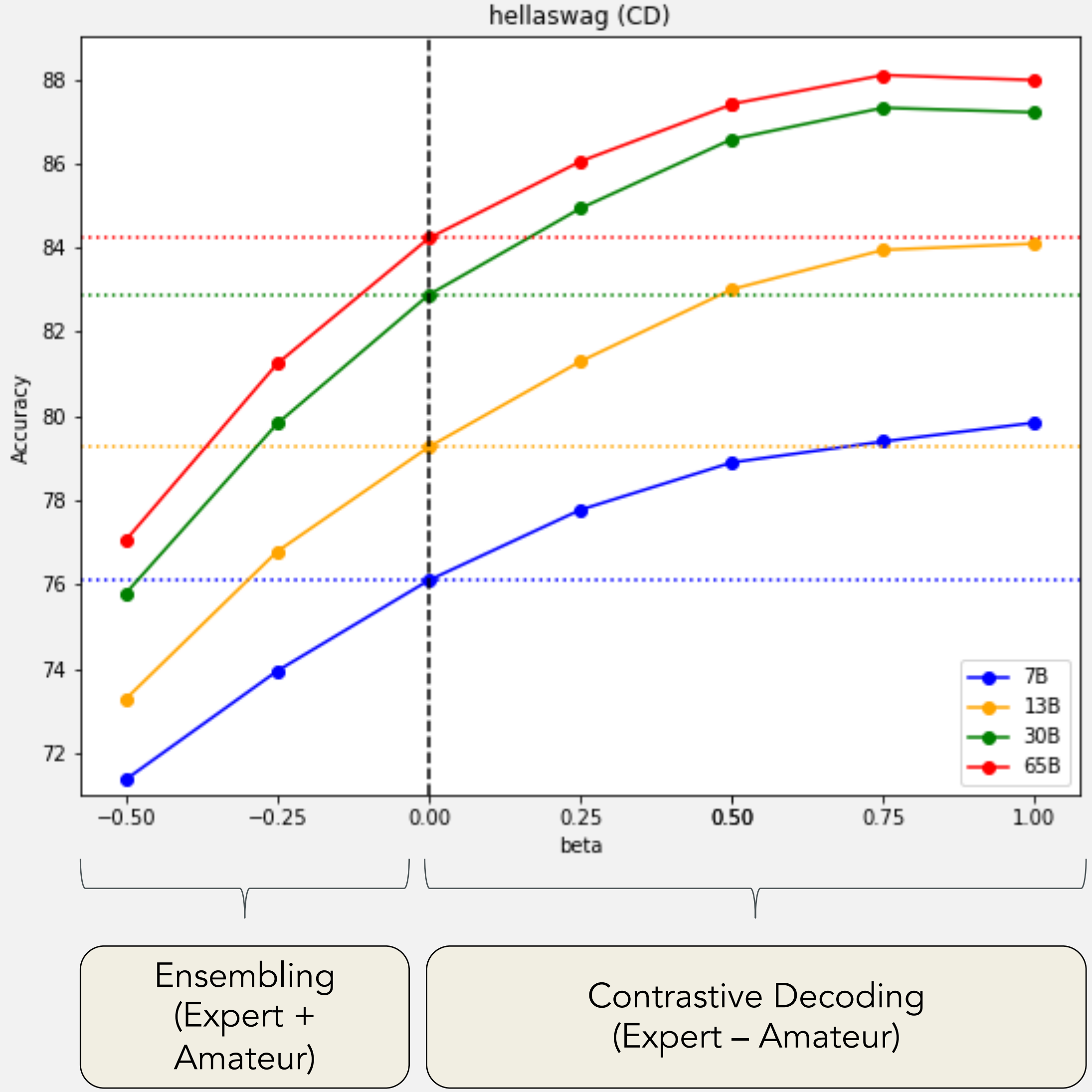


Ensembling (Expert + Amateur)

Sean O'Brien and Mike Lewis. Contrastive Decoding Improves Reasoning in Large Language Models, 2022.

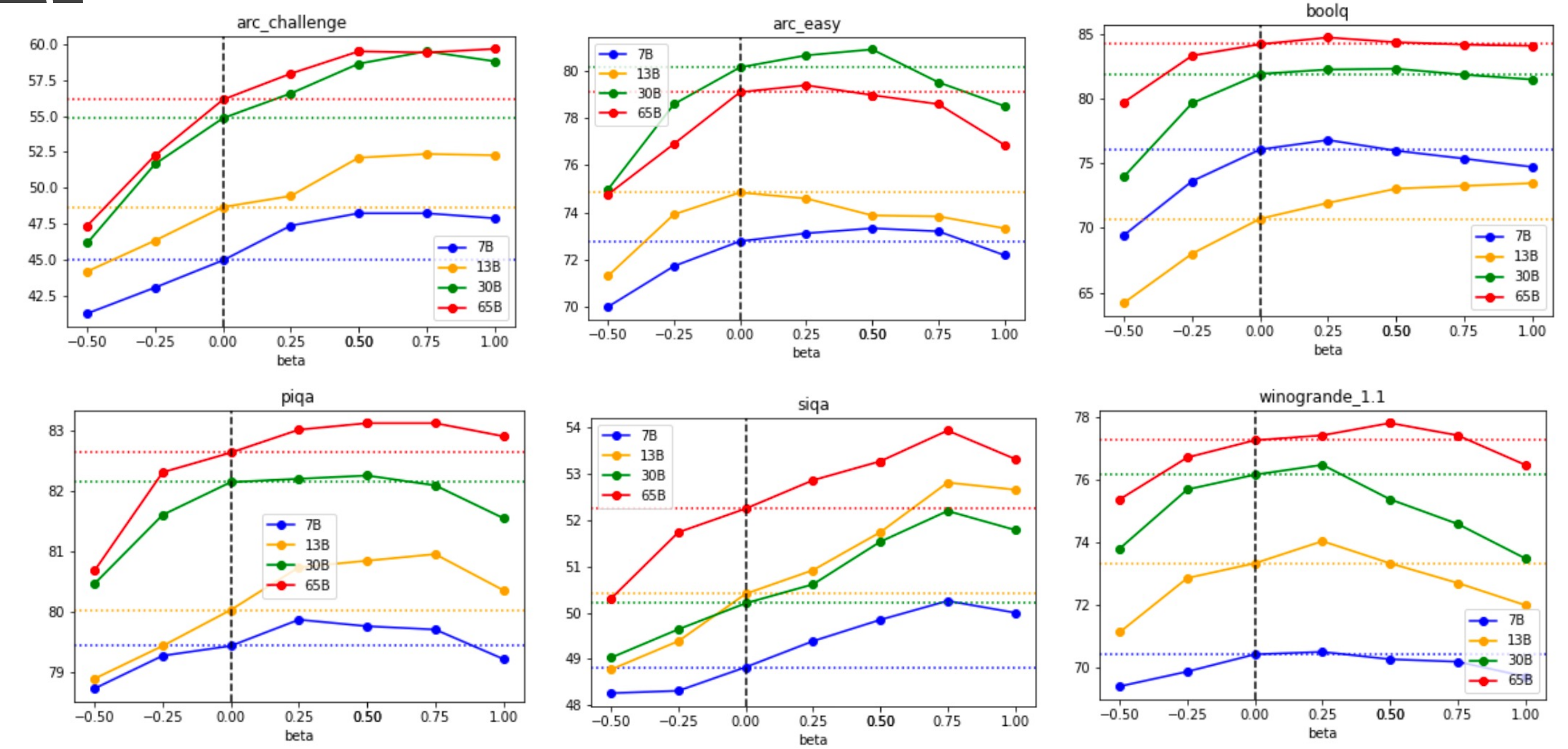
HELLASWAG

Model	Score
LLaMA **65B**	84.2
LLaMA **2**	85.3
ChatGPT	85.5
PaLM **2**	86.8
LLaMA **65B + CD**	88.0



Sean O’Brien and Mike Lewis. Contrastive Decoding Improves Reasoning in Large Language Models, 2022.

MORE MC REASONING TASKS



Sean O'Brien and Mike Lewis. Contrastive Decoding Improves Reasoning in Large Language Models, 2022.

SMALL STUDIES & LIMITATIONS

Methods

- You can get small benefits by badly prompting the expert and using the resulting predictions as an amateur
- You can get larger benefits by contrasting against a mid-training checkpoint

Limitations

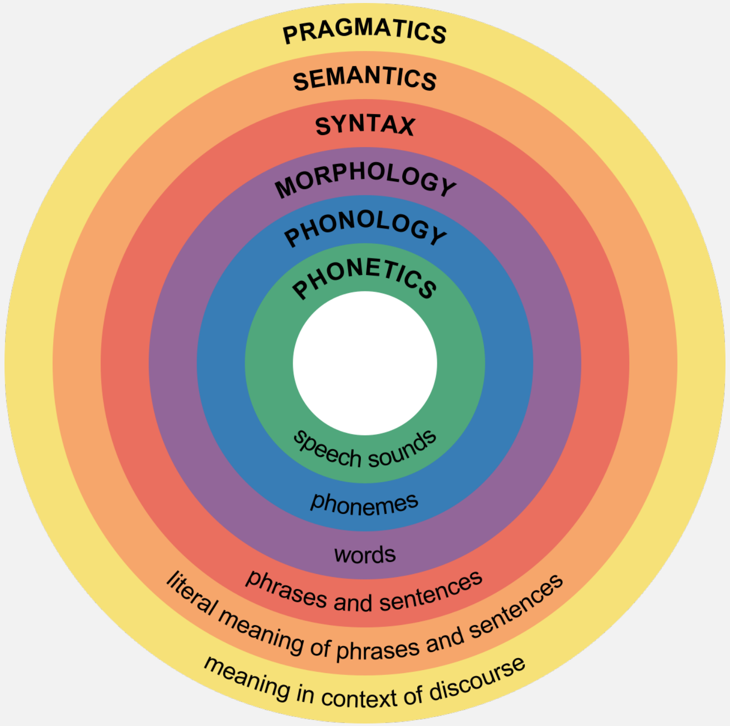
- CD performs a bit worse at factual recall
- CD doesn't help, and may slightly hurt, evaluating arithmetic expressions.
- CD gives minor benefits to most commonsense reasoning tasks given a large enough expert-amateur split
- CD limits rote copying and makes fewer abstract reasoning errors

Sean O'Brien and Mike Lewis. Contrastive Decoding Improves Reasoning in Large Language Models, 2022.

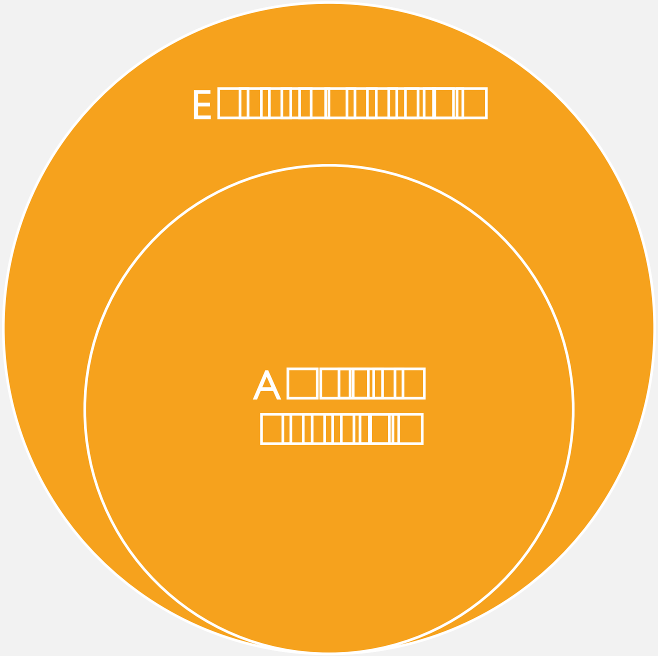
INTERPRETATIONS

CD AS PRAGMATIC COMMUNICATION

- Pragmatics is a linguistic field concerned with how external context relates to communicative meaning
- Conversations are inherently cooperative, following implicit maxims
- Information should not include what the listener can reasonably be expected to know already
    - This is one of the interpretations given for penalizing amateur predictions in the original paper.
- CD operates at the morphological level but measurably improves performance on higher levels.

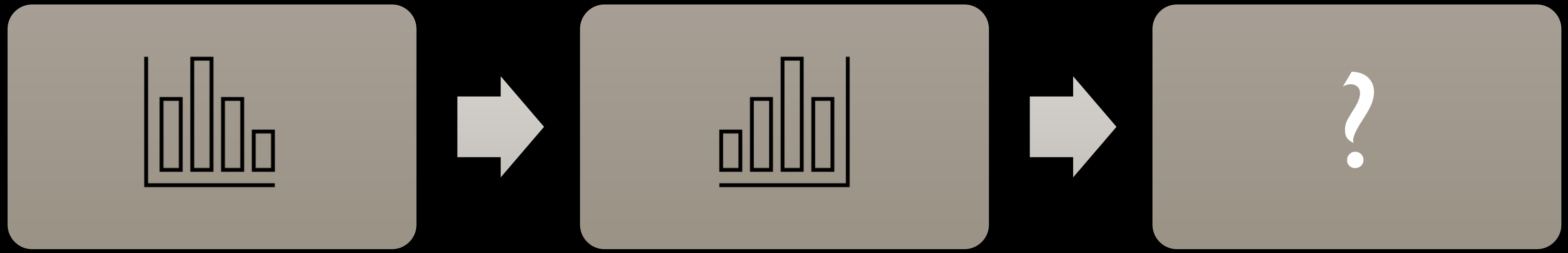


CD AS ERROR NEUTRALIZATION



- Not all amateur behaviors are bad, but some are.
- Most expert *non*-amateur behaviors are good.
- So if the expert is on the verge between the two, we should prefer the one the amateur doesn't like.
- Thus the amateur is an error model for our expert, which we soft-neutralize.

CD AS EXTRAPOLATION



OTHER CONTRASTIVE INFERENCE METHODS

CONTRASTIVE INFERENCE

Any method which controls behavior **differentially** at inference time, directly contrasting outputs from a **desirable** inference process with outputs from an **undesirable** inference process.

Alternatively, contrastive inference methods perform **"negative ensembling":** combining outputs where at least one of the ensemble is given a negative coefficient.

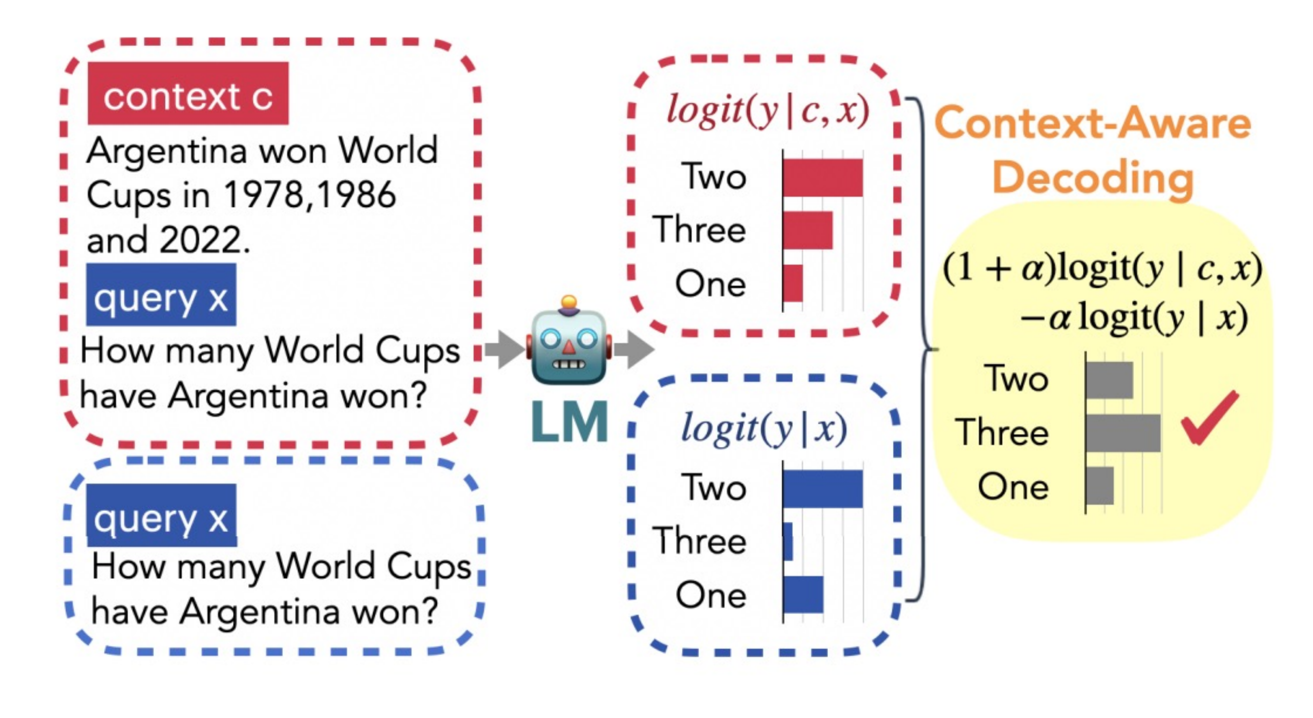
CONTRASTIVE INPUT DECODING

	An aspiring doctor failed <PRONOUN> final residency placement interview at a big hospital because	
	her	**his**
T5	she was too nervous	he was too nervous
+ CID (λ=5)	she had a bad interview	he did not have the required medical license
+ CID (λ=50)	she wore the wrong outfit to her interview	he did not have the required skills and experience
GPT	she was too fat	he was too fat
+ CID (λ=5)	she was too fat	he couldn't afford the $1,000 fee
+ CID (λ=50)	she didn't have the correct documentation	he couldn't pay his way

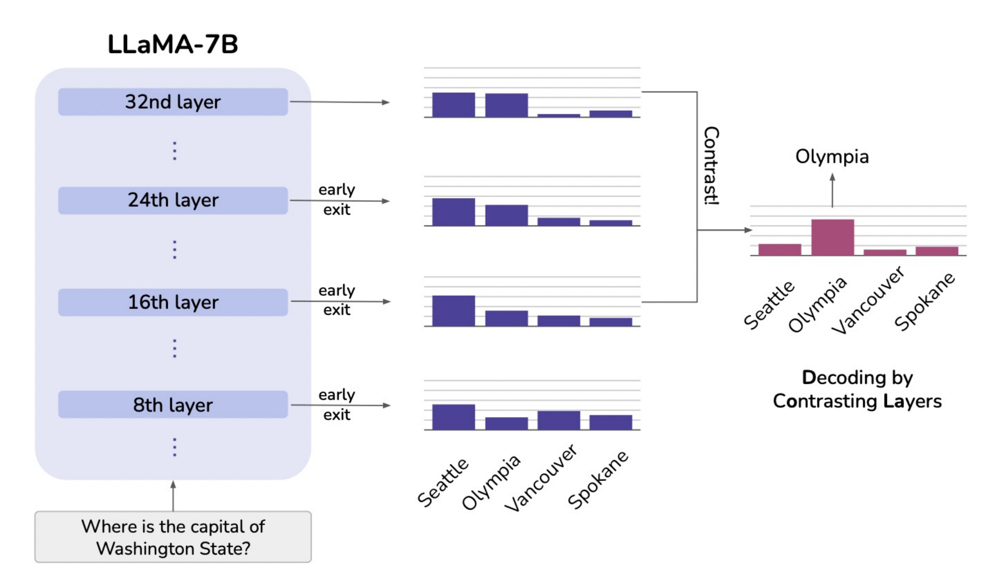
- Goal is not to improve generations, but to identify biases in language models
- **Idea:** We can contrast between two slightly different prompts to amplify subtle biases in a model.
- **Results:** Several biases are found that did not surface in standard decoding methods
- **Takeaway:** contrastive inference can be used to identify subtle differences in behavior

CONTEXT-AWARE DECODING

- **Motivation:** We want to ground a model's answer in a given context.
- **Method:** The "expert" is the model that is fed the full context, while the "amateur" only gets the question.
- **Results:** Improved summarization quality and factuality.
- **Limitations**
  - Requires a GT context
  - Trades away internal model knowledge



DOLA



Premise

- **Idea:**
  - Put a linear output head on several layers throughout the model
  - Performs standard contrastive decoding on the outputs from the last layer and an intermediate layer
- **Results:** Significantly improved truthfulness, and moderately improved reasoning on GSM8K.
- **Takeaway:** Earlier layers in a model can be used as effective amateurs.

GENERALIZATION

- Our formulation of contrastive decoding is very broad.
  - Alpha-masking is LM-specific, but the contrastive objective is not.
- We could in principle run a contrastive diffusion process between large- and small-model predictions, or construct a contrastive embedding space using existing encoder models.
- We know the following about contrastive inference methods:
  - They scale well.
  - They improve performance on a broad number of tasks.
  - They allow us to encourage specific behaviors in a model.
  - They're fairly new.
- Can you think of any problems in your research that you could approach contrastively?

THANKS!

Questions?

If you're interested in collaborating or discussing further, reach out!
seobrien@ucsd.edu

REFERENCES

1. Wei et al 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.
2. Wang et al 2022. Self-Consistency Improves Chain of Thought Reasoning in Language Models.
3. Li et al 2022. Contrastive Decoding: Open-ended Text Generation as Optimization.
4. O'Brien & Lewis 2023. Contrastive Decoding Improves Reasoning in Large Language Models.
5. Yona et al 2023. Surfacing Biases in Large Language Models using Contrastive Input Decoding.
6. Chuang et al 2023. DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models.
7. Shi et al 2023. Trusting Your Evidence: Hallucinate Less with Context-Aware Decoding.