SFU NatLangLab

CMPT 713: Natural Language Processing

Dialogue and large language models

Spring 2023

2023-03-23

Dialogue

What's a Dialogue System?

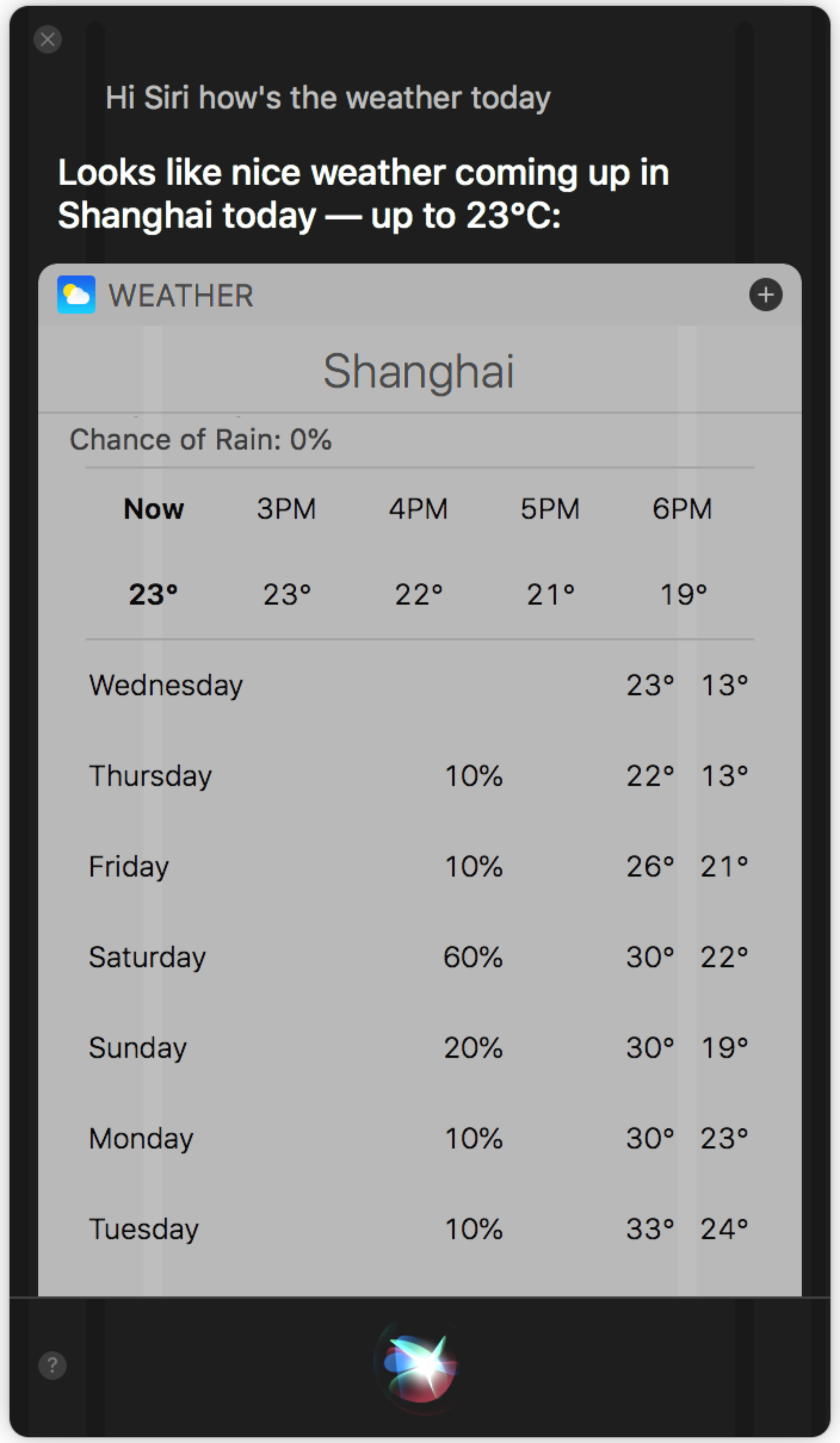
Dialog Systems are **HOT** 🔥. — Did you use it?

How ChatGPT Kicked Off an A.I. Arms Race

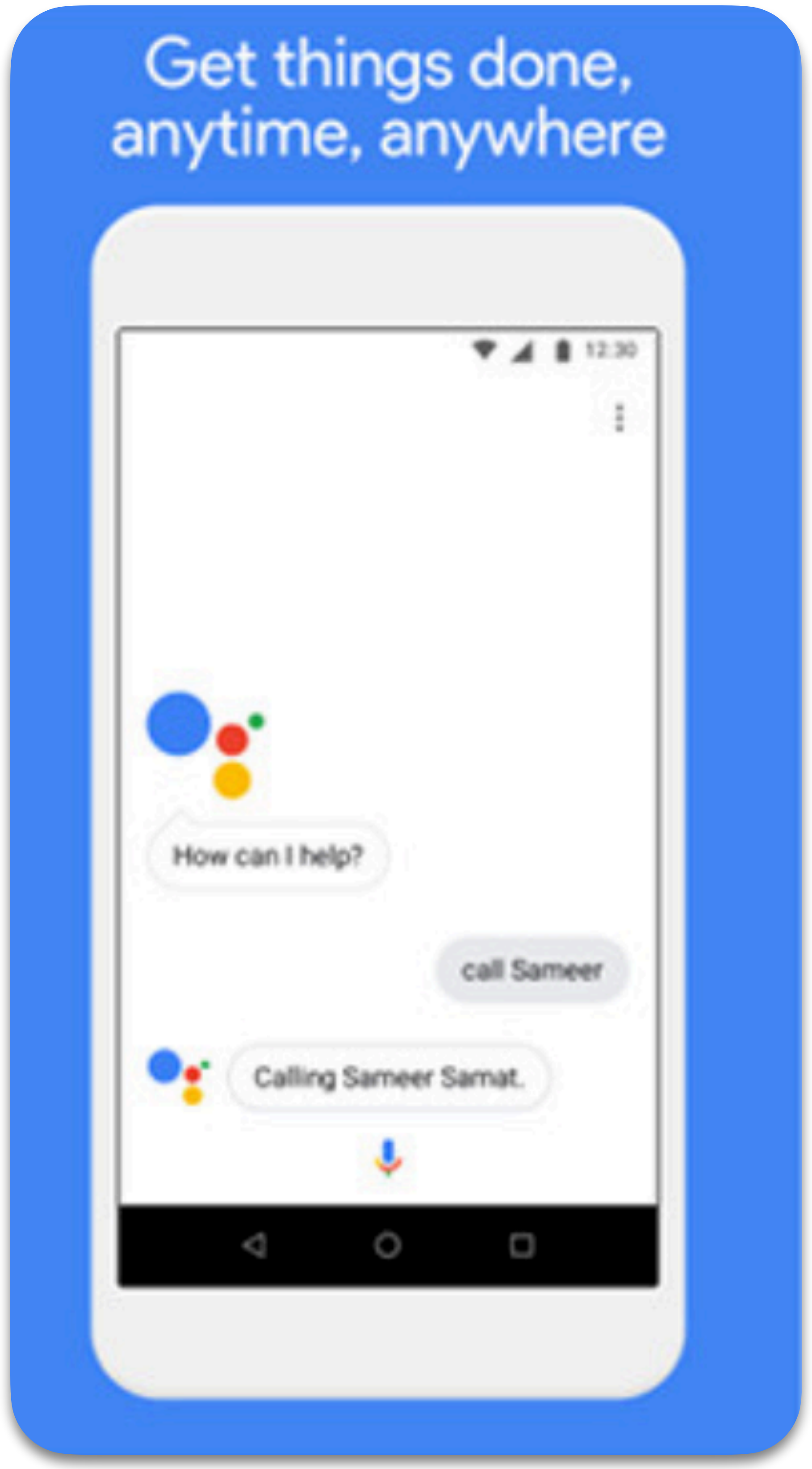
Feb, 2023

Conversational agents

Siri



Apple



Google

Amazon

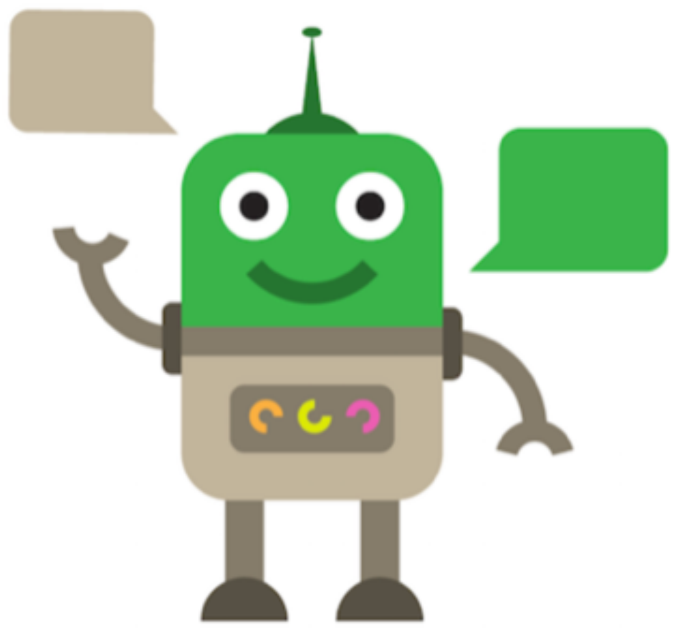
alexa

Microsoft

Hey Cortana

Two kinds of conversational agents

- **Chatbots**
  - Mimic informal human chatting
  - For fun, or even for therapy
- **Task-oriented dialog agents**
  - Interfaces to personal assistants
  - Cars, robots, appliances
  - Booking flights or restaurants

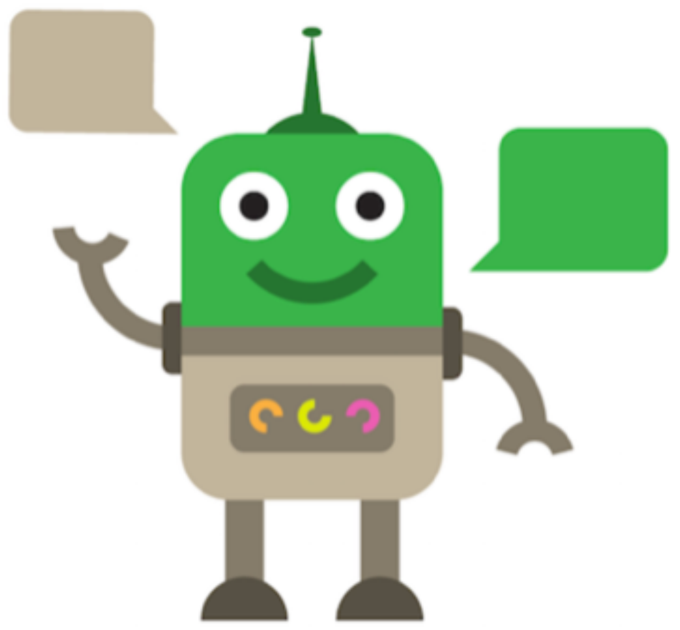


Dialogue Systems

Chatbot architectures

- **Rule-based**
  - Pattern-action rules (ELIZA)
  - + A mental model (PARRY):
    - First system to pass the Turing test!
- **Corpus-based (data-driven)**
  - Information Retrieval (XiaoIce)
  - Neural encoder-decoder (BlenderBot)

- **Chatbots**
  - Mimic informal human chatting
  - For entertainment



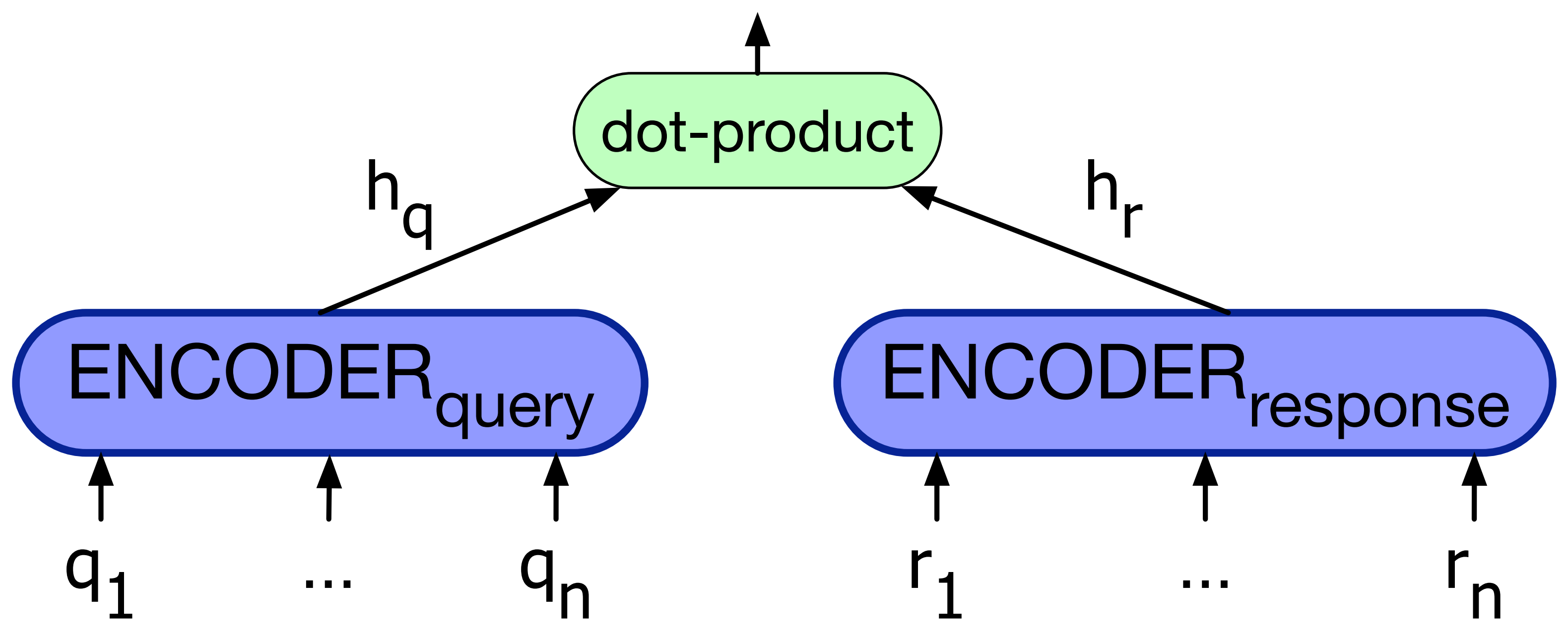
Dialogue Systems

Chatbot: Corpus-based methods

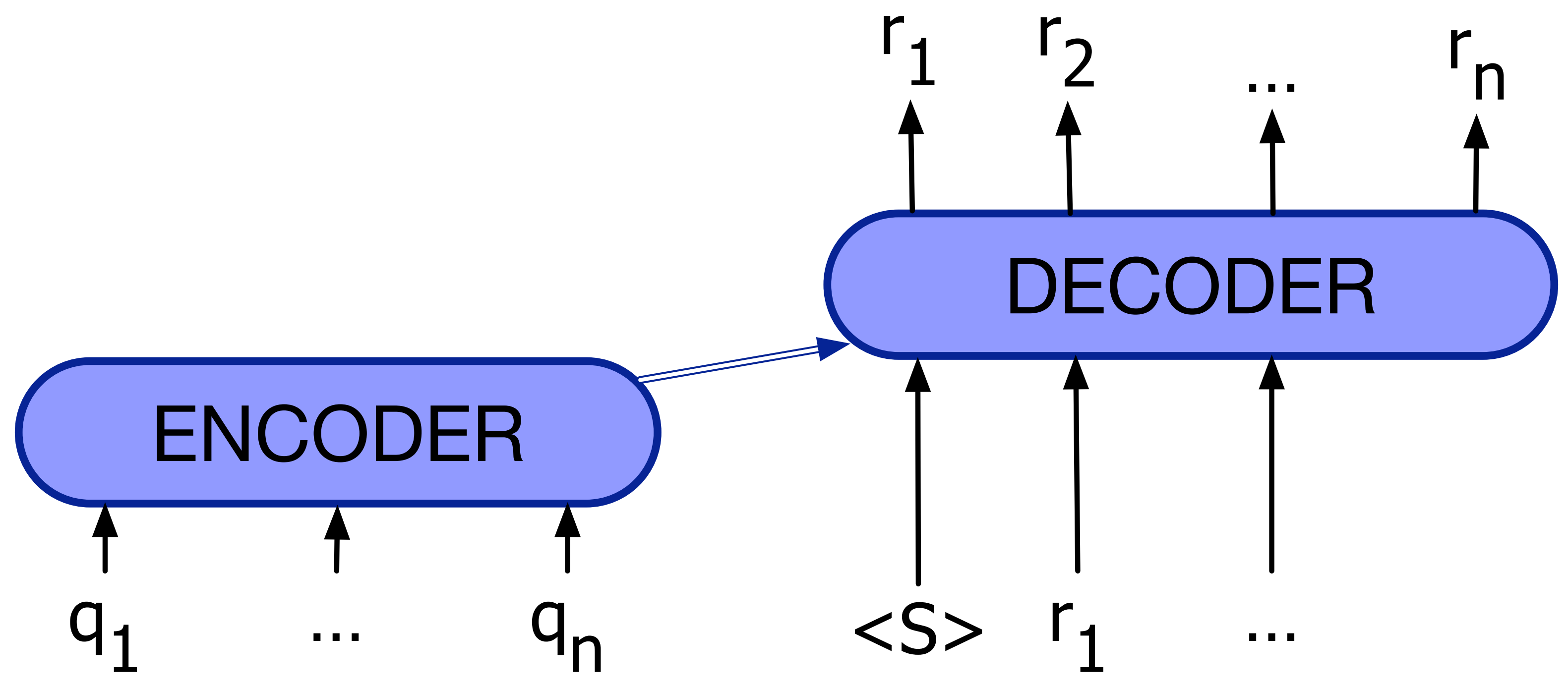
Response by retrieval
- Use information retrieval to grab a response (that is appropriate to the context) from some corpus

Response by generation
- Use a language model or encoder-decoder to generate the response given the dialogue context

Both methods require large corpora to perform well!
- Modern corpus-based chatbots are very data-intensive
- Typically require hundreds of millions or billions of words
- Corpus come from transcripts of telephone conversations, movie dialogues, conversations on social media, crowd-sourced data.



(a) Response by Retrieval



(b) Response by Generation

Corpus-based method (Response by retrieval)

Return the response to the most similar turn.

Neural IR method

Given user query q and a conversational corpus C. ← Training corpus
Find the response r in C that has an encoding that is most similar to the encoding of q,.

Bi-encoder: two separate encoders

$$h_q = \textbf{BERT}_Q(q)[\textbf{CLS}]$$
$$h_r = \textbf{BERT}_R(r)[\textbf{CLS}]$$

$$\textbf{response}(q, C) = \underset{r \in C}{\arg\max}(h_q \cdot h_r)$$

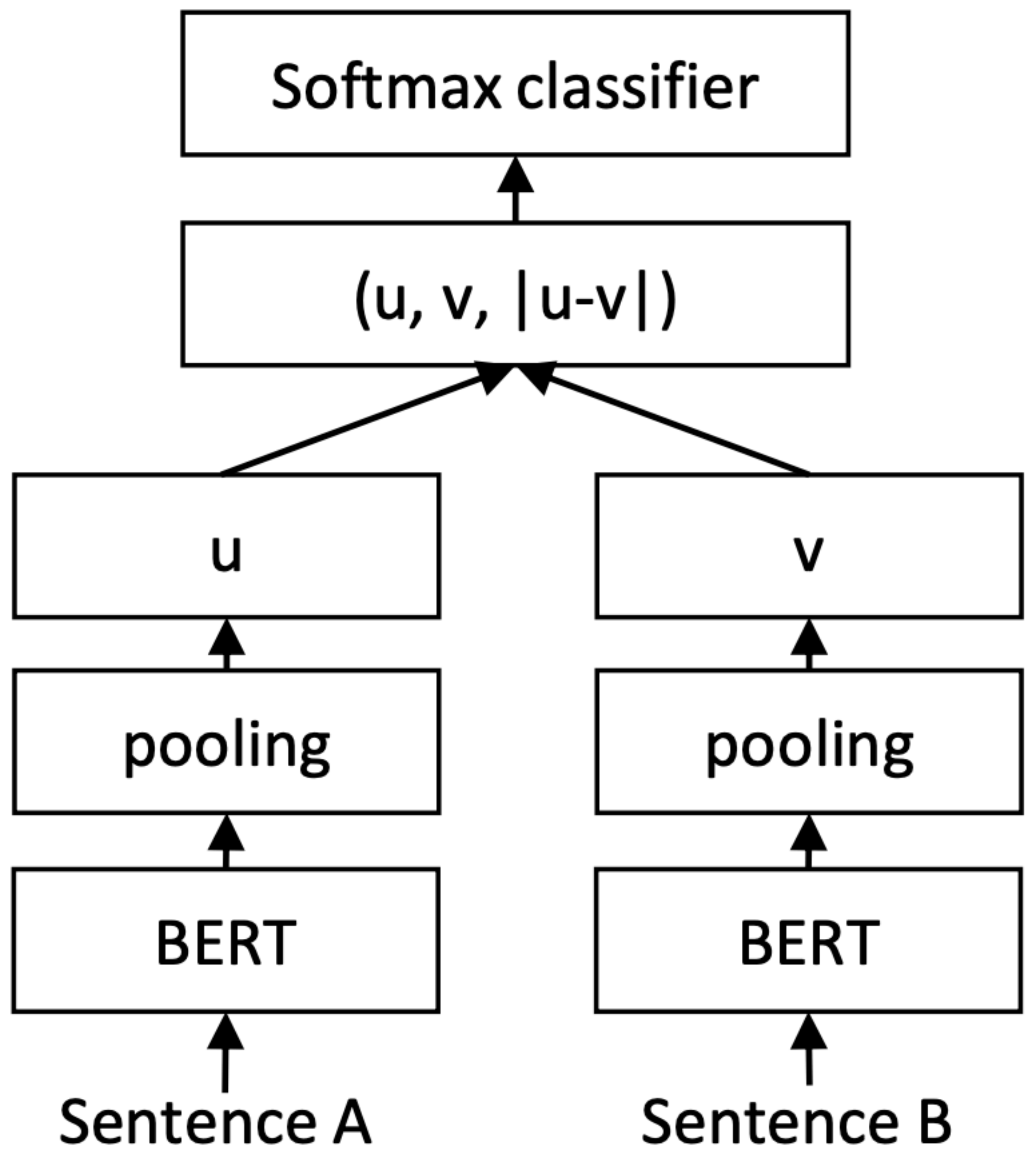
Similarity from neural models

Can also have more sophisticated neural architectures

Chatbot

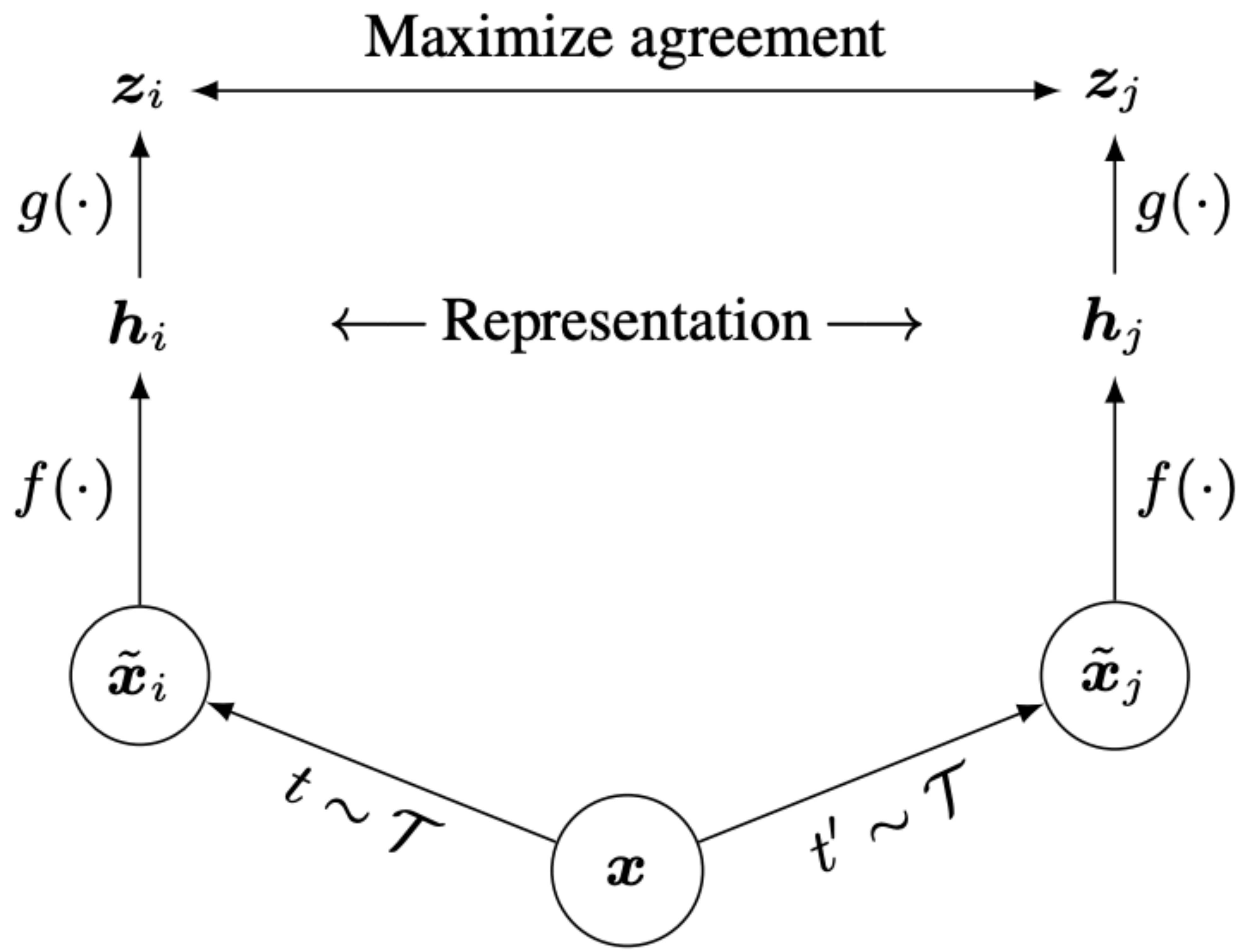
Learning sentence embeddings

- Train sentence embedder on supervised data
- Pooling options:
  - CLS token, MEAN, MAX



[Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks](),Reimers and Gurevych, EMNLP 2019

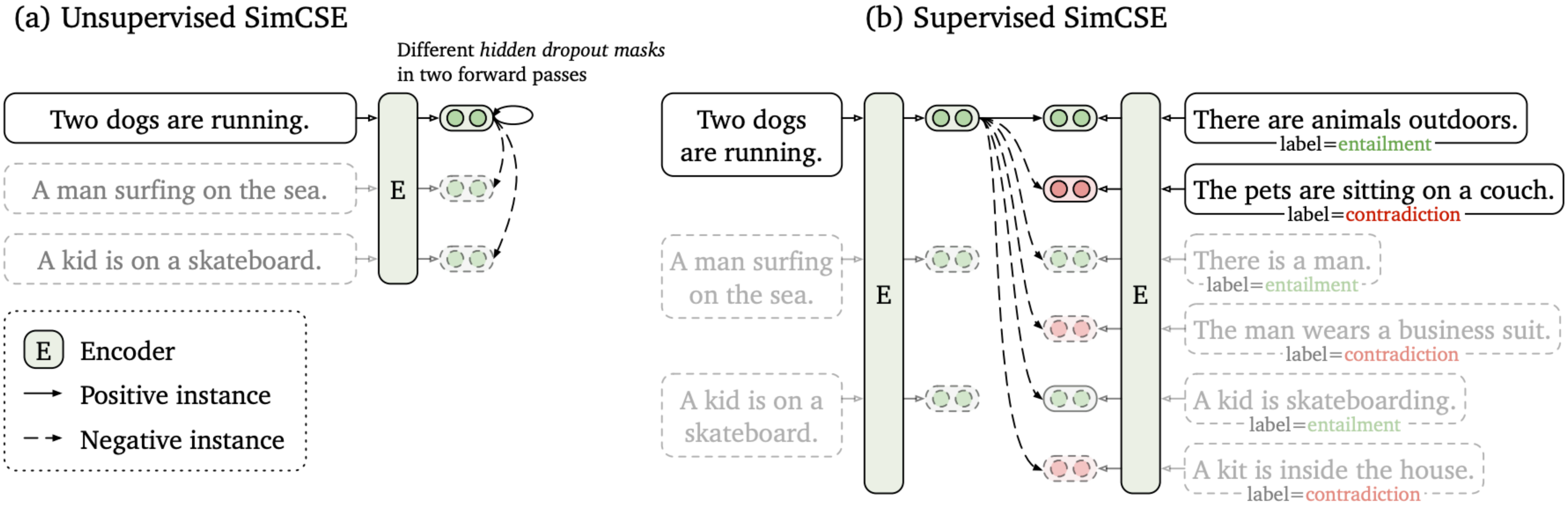
- What about unsupervised contrastive learning?



SimCLR [Chen et al. ICML 2020]

Learning sentence embeddings

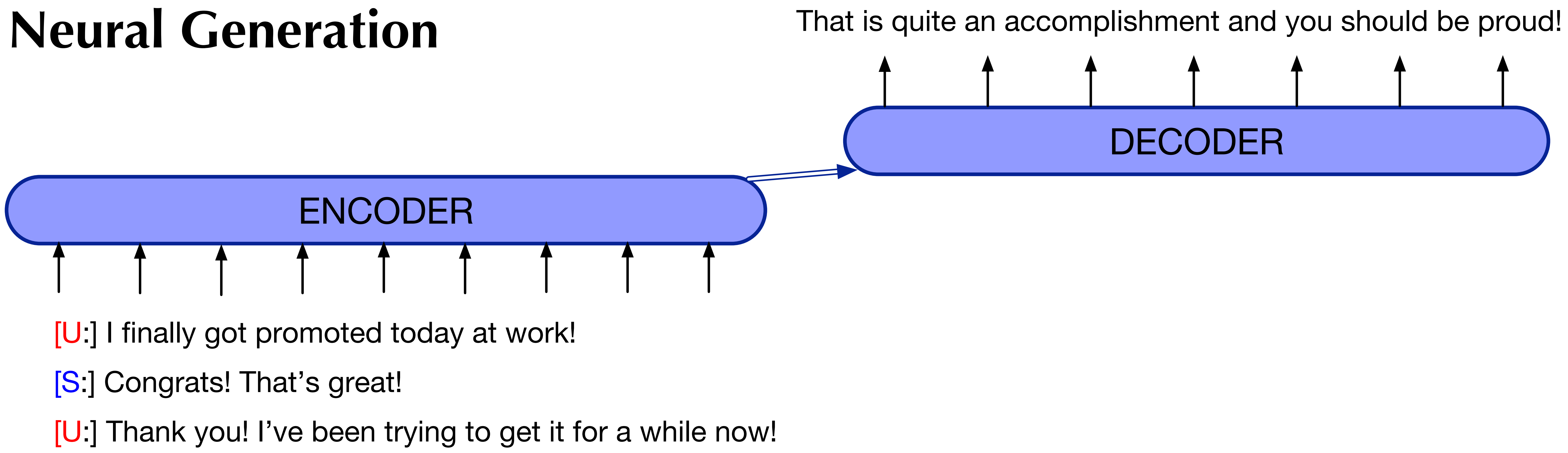
Use dropout for unsupervised contrastive learning of sentence embeddings



SimCSE: Simple Contrastive Learning of Sentence Embeddings, Gao et al, EMNLP 2021

Corpus-based method (Response by generation)

An encoder decoder model for neural response generation in dialogue.



Can train just on conversation data or fine-tune a large language model on conversational data

Fine tuning: Chirpy Cardinal System (Paranjape et al., 2020)

- Fine-tunes GPT-2 on EmpatheticDialogues dataset (Rashkin et al, 2019)
- Note that this is a Decoder only model

Chatbot: Seq2Seq models

Repetitive

A: Where are you going?
B: I'm going to the restroom.
A: See you later.
B: See you later.
A: See you later.
B: See you later.

Sample and Rank

1. Sample N candidate
2. Rank candidate and select best one

Sampling

Randomly sample words from distribution at each time step t

- **Basic/pure sampling**: sample from $P_t(w)$ directly
  - Can get some very bad samples
  - No control
- **Top-n sampling**: sample from P_t truncated to top n words
  - Increase n to get more diverse/risky output
  - Decrease n to get more generic/safe output
- **Top-p sampling**: sample from P_t restricted to top p proportion of words
  - Better when probability distribution is spread
- **Temperature based sampling**:
  - Increase τ to get more diverse/risky output (P_t is more uniform)
  - Decrease τ to get more generic/safe output (P_t is more spiky)

$$P_t(w) = \frac{\exp(s_w/\tau)}{\sum_{w' \in V} \exp(s_{w'}/\tau)}$$

(adapted from slides: Stanford CS224N, Chris Manning)

Task-based dialogue agents

- **"Task-based"** or **"goal-based"** dialogue agents
  - Systems that have the goal of helping a user solve a task
    - Setting a timer
    - Making a travel reservation
    - Playing a song
    - Buying a product
- Need to incorporate task-specific knowledge
  - **Frames** with **slots** and **values**
  - A knowledge structure representing user **intentions**

Task-based dialogue agents

- **Frame**
  - Contains a set of **slots**, to be filled with information of a given **type**.
  - Each associated with a **question** to the user

Slot	Type	Question
ORIGIN	city	What city are you leaving from?
DEST	city	Where are you going?
DEP DATE	date	What day would you like to leave?
DEP TIME	time	What time would you like to leave?
AIRLINE	Airline	What is your preferred airline?

How to detect frames and fill in dialog slots?

Natural language understanding

"Show me morning flights from
Boston to San Francisco on Tuesday"

Before filling in the dialog slots:

- System must detect which slot of which frame user is filling
- And switch dialogue control to that frame.

How to detect frames and fill in dialog slots?

Natural language understanding

"Show me morning flights from Boston to San Francisco on Tuesday"

Step#1: domain classification

DOMAIN: AIR-TRAVEL

Classification

How to detect frames and fill in dialog slots?

Natural language understanding

"Show me morning flights from
Boston to San Francisco on Tuesday"

Identify the frame to use

Step#1: domain classification DOMAIN: AIR-TRAVEL

Step#2: intent determination INTENT: SHOW-FLIGHTS

Classification

How to detect frames and fill in dialog slots?

Natural language understanding

"Show me morning flights from
Boston to San Francisco on Tuesday"

Step#1: domain classification	DOMAIN: AIR-TRAVEL
Step#2: intent determination	INTENT: SHOW-FLIGHTS
Step#3: slot filling	ORIGIN-CITY: Boston ORIGIN-DATE: Tuesday ORIGIN-TIME: morning DEST-CITY: San Francisco

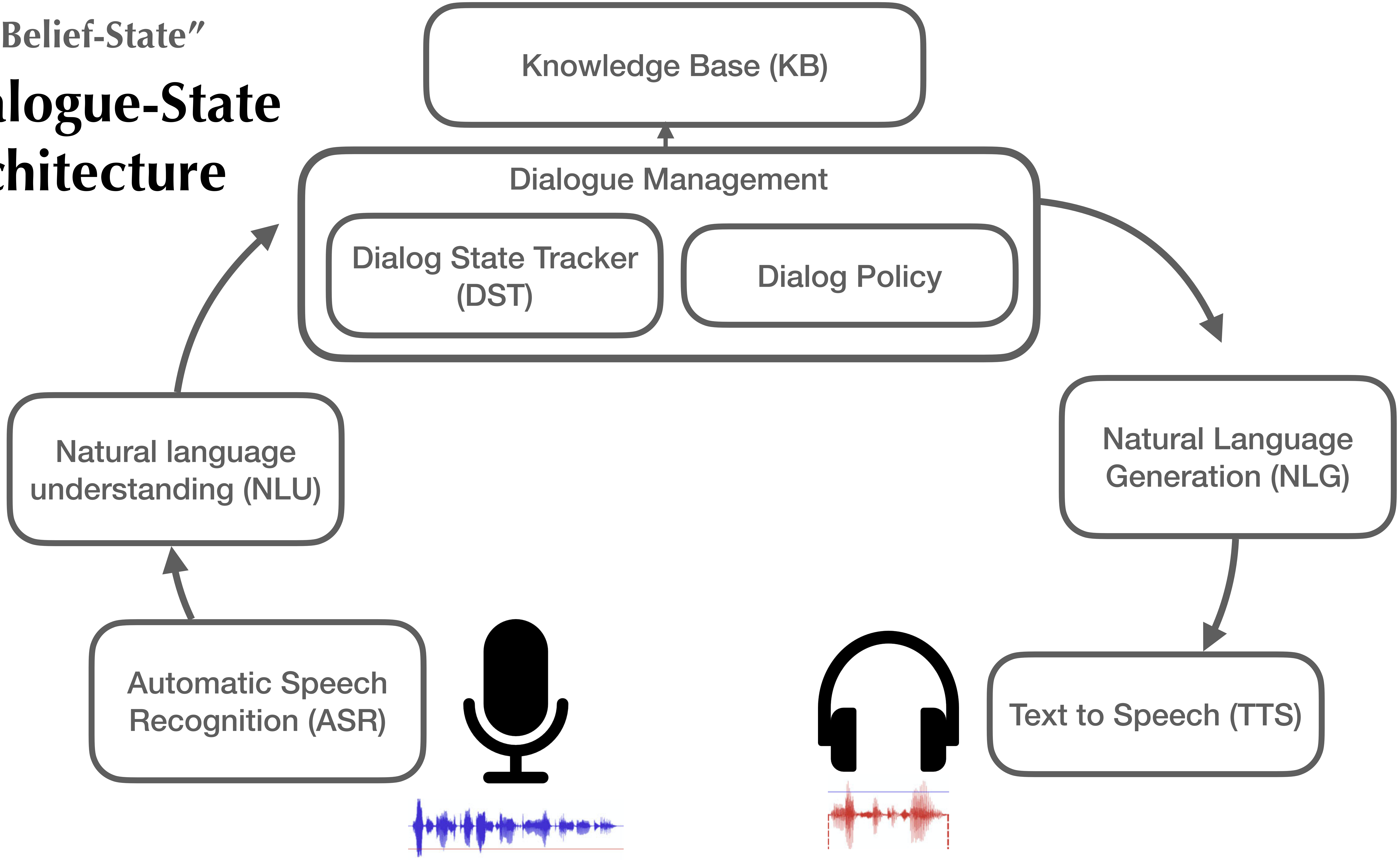
Rule-based

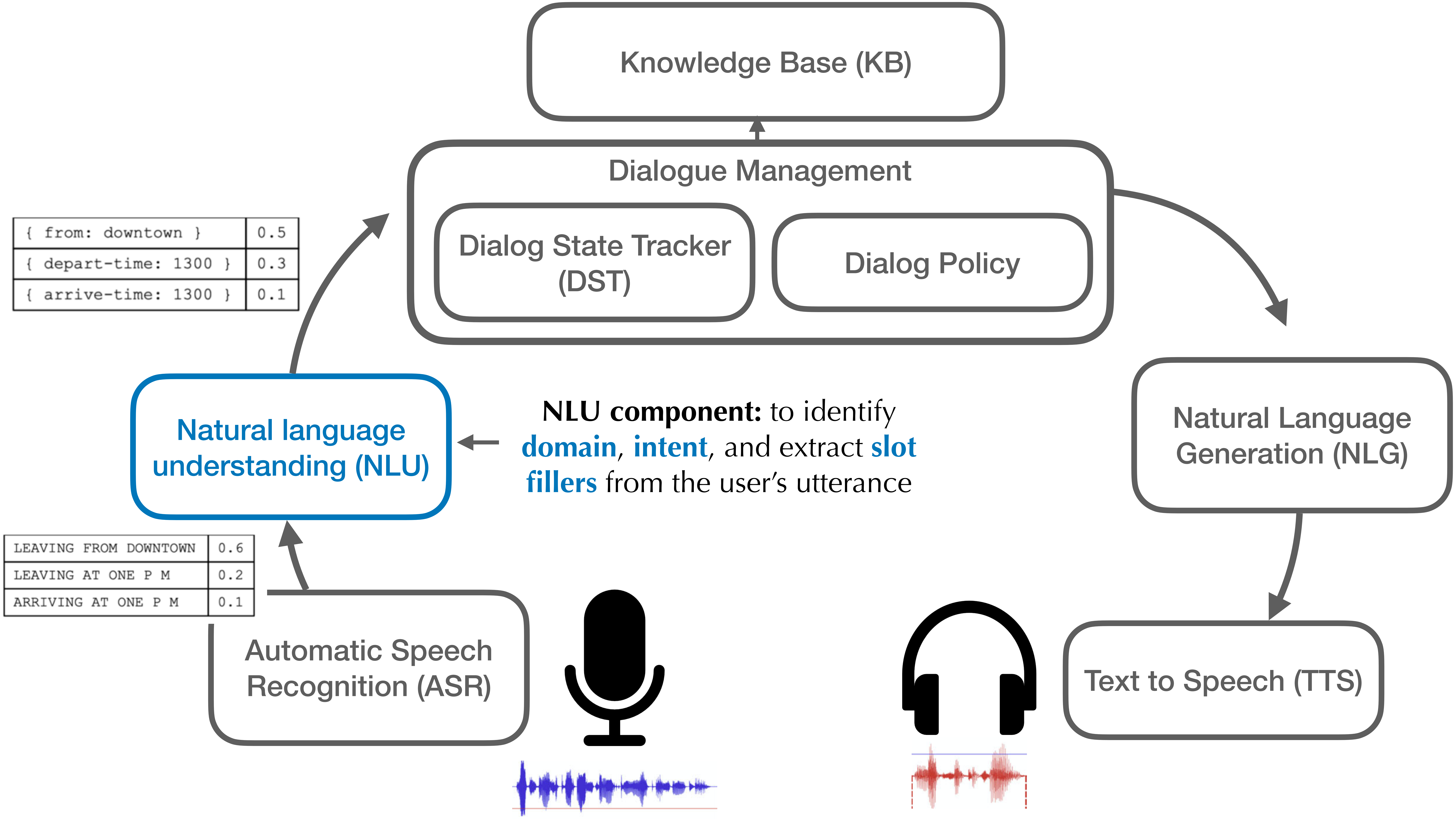
or

Sequence tagging

“Belief-State”

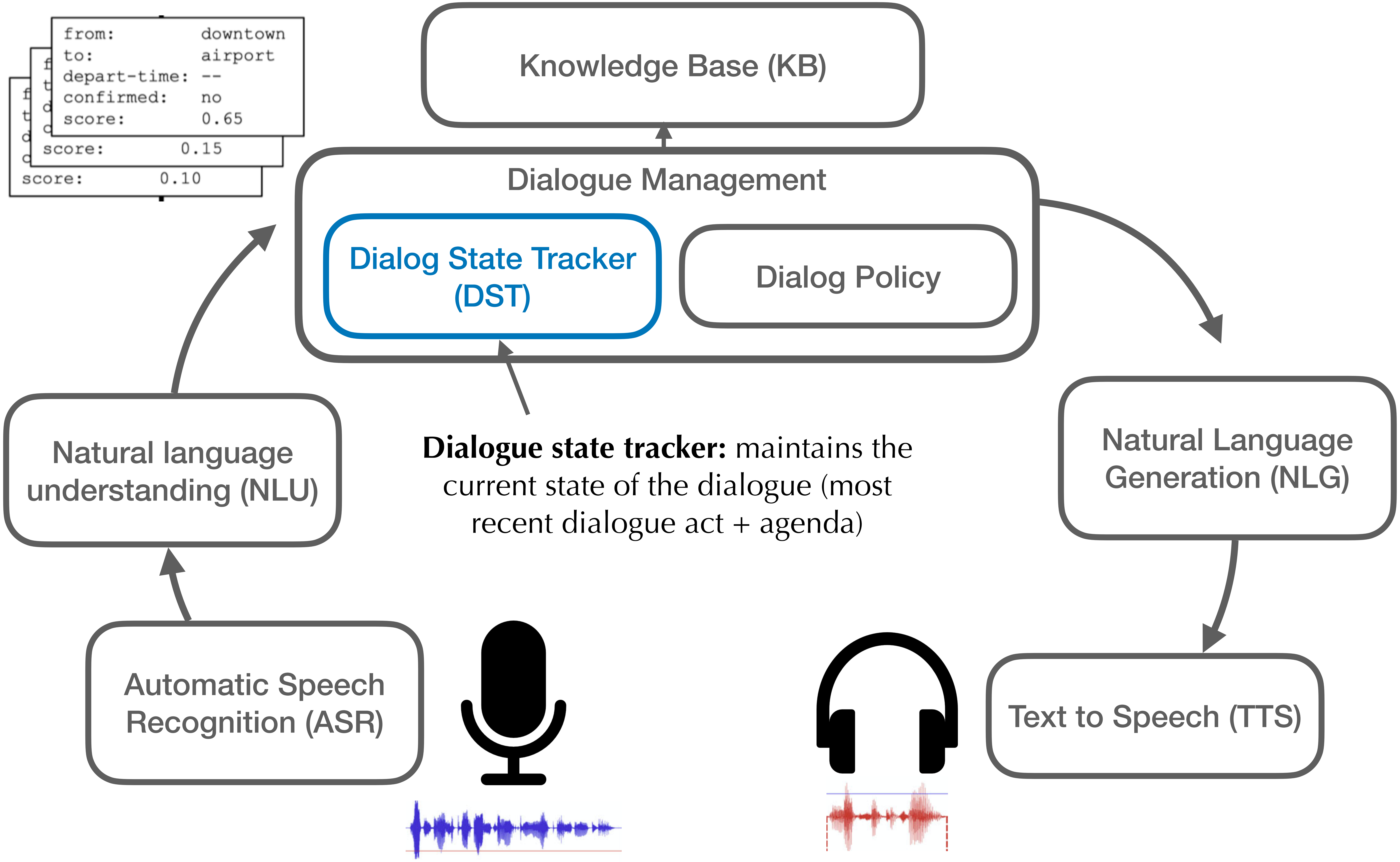
Dialogue-State Architecture

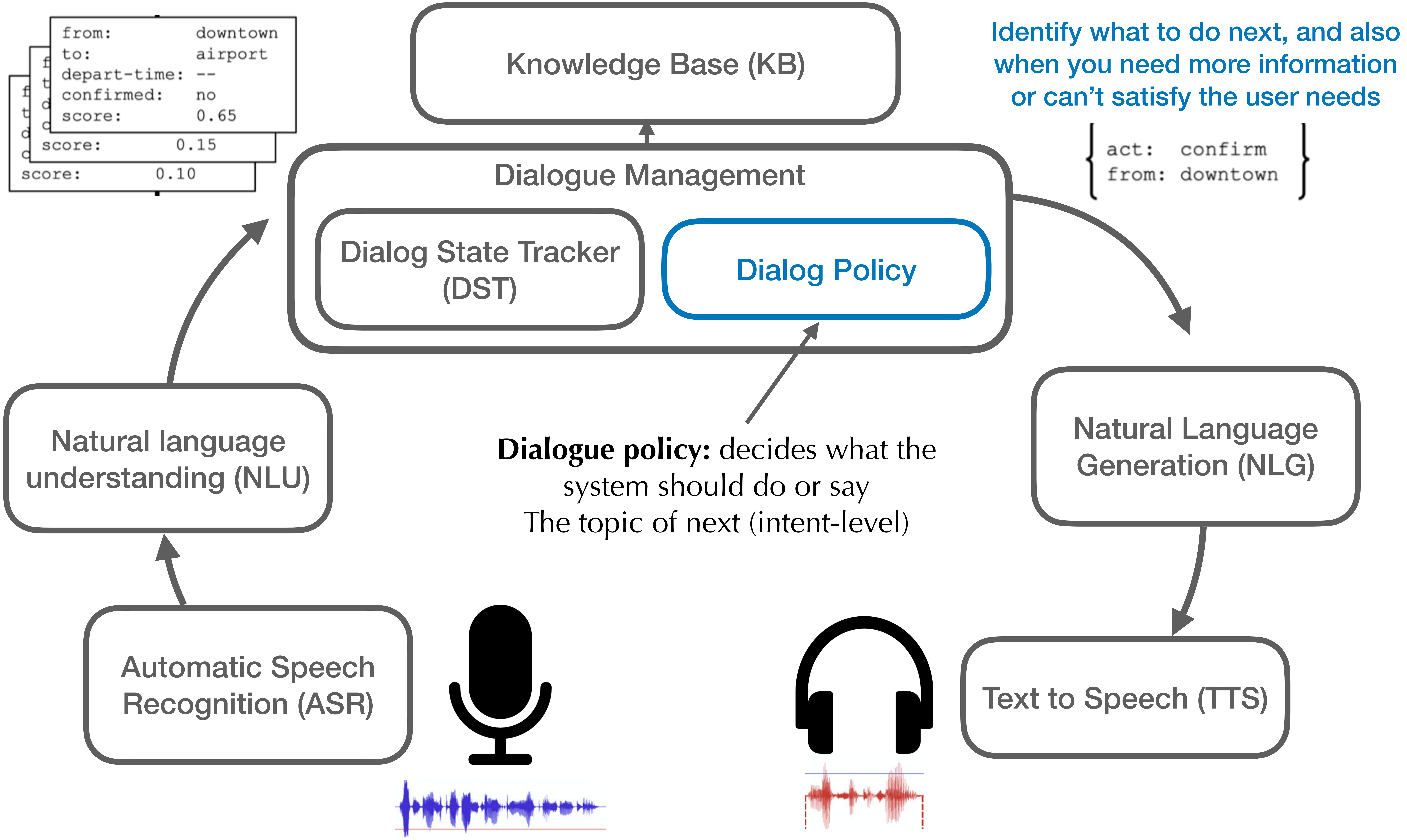


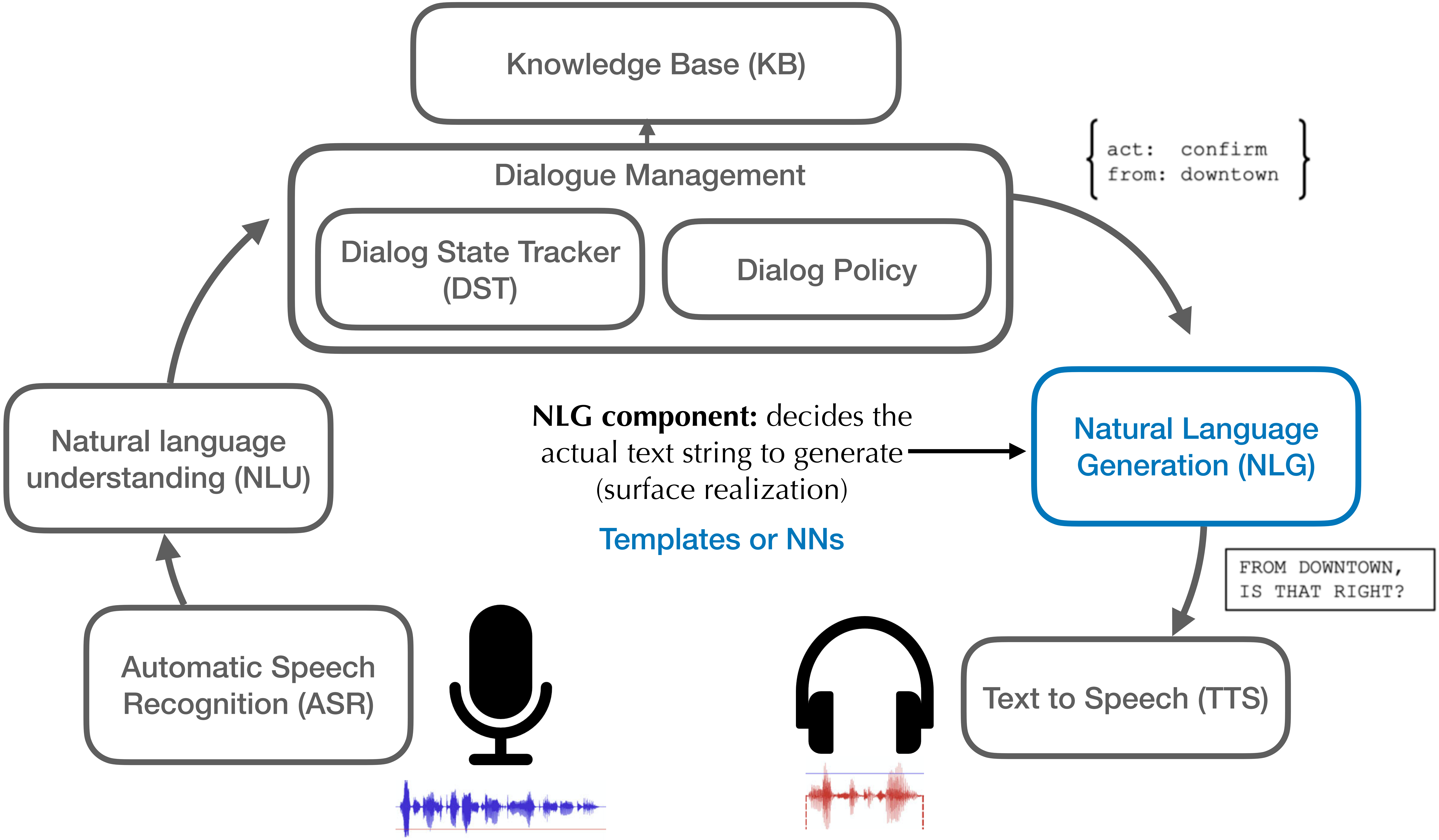


{ from: downtown }	0.5
{ depart-time: 1300 }	0.3
{ arrive-time: 1300 }	0.1

LEAVING FROM DOWNTOWN	0.6
LEAVING AT ONE P M	0.2
ARRIVING AT ONE P M	0.1


from: downtown
to: airport
depart-time: --
confirmed: no
score: 0.65
score: 0.15
score: 0.10
Knowledge Base (KB)
Dialogue Management
Dialog State Tracker (DST)
Dialog Policy
Natural language understanding (NLU)
Natural Language Generation (NLG)
Automatic Speech Recognition (ASR)
Text to Speech (TTS)
Dialogue state tracker: maintains the current state of the dialogue (most recent dialogue act + agenda)


from: downtown
to: airport
depart-time: --
confirmed: no
score: 0.65
score: 0.15
score: 0.10
Knowledge Base (KB)
Identify what to do next, and also when you need more information or can't satisfy the user needs
act: confirm
from: downtown
Dialogue Management
Dialog State Tracker (DST)
Dialog Policy
Natural language understanding (NLU)
Dialogue policy: decides what the system should do or say
The topic of next (intent-level)
Natural Language Generation (NLG)
Automatic Speech Recognition (ASR)
Text to Speech (TTS)



Knowledge Base (KB)
Dialogue Management
Dialog State Tracker (DST)
Dialog Policy
act: confirm
from: downtown
Natural language understanding (NLU)
NLG component: decides the actual text string to generate (surface realization)
Templates or NNs
Natural Language Generation (NLG)
FROM DOWNTOWN, IS THAT RIGHT?
Automatic Speech Recognition (ASR)
Text to Speech (TTS)

Dialogue System Evaluation

Chatbots:

- Mostly human evaluation
- Automatic evaluation is a challenge
  - Use automatic metrics to capture specific aspect (diversity, length of conversation)
  - Adversarial evaluation

Task-based dialogue agents

- Automatic metrics to evaluate task performance
- Can also use human evaluation

Automatic evaluation metrics

Content overlap metrics:

- Word (n-gram) overlap: BLEU, ROUGE, METEOR, CIDEr
- Structured overlap: PYRAMID, SPICE, SPIDER

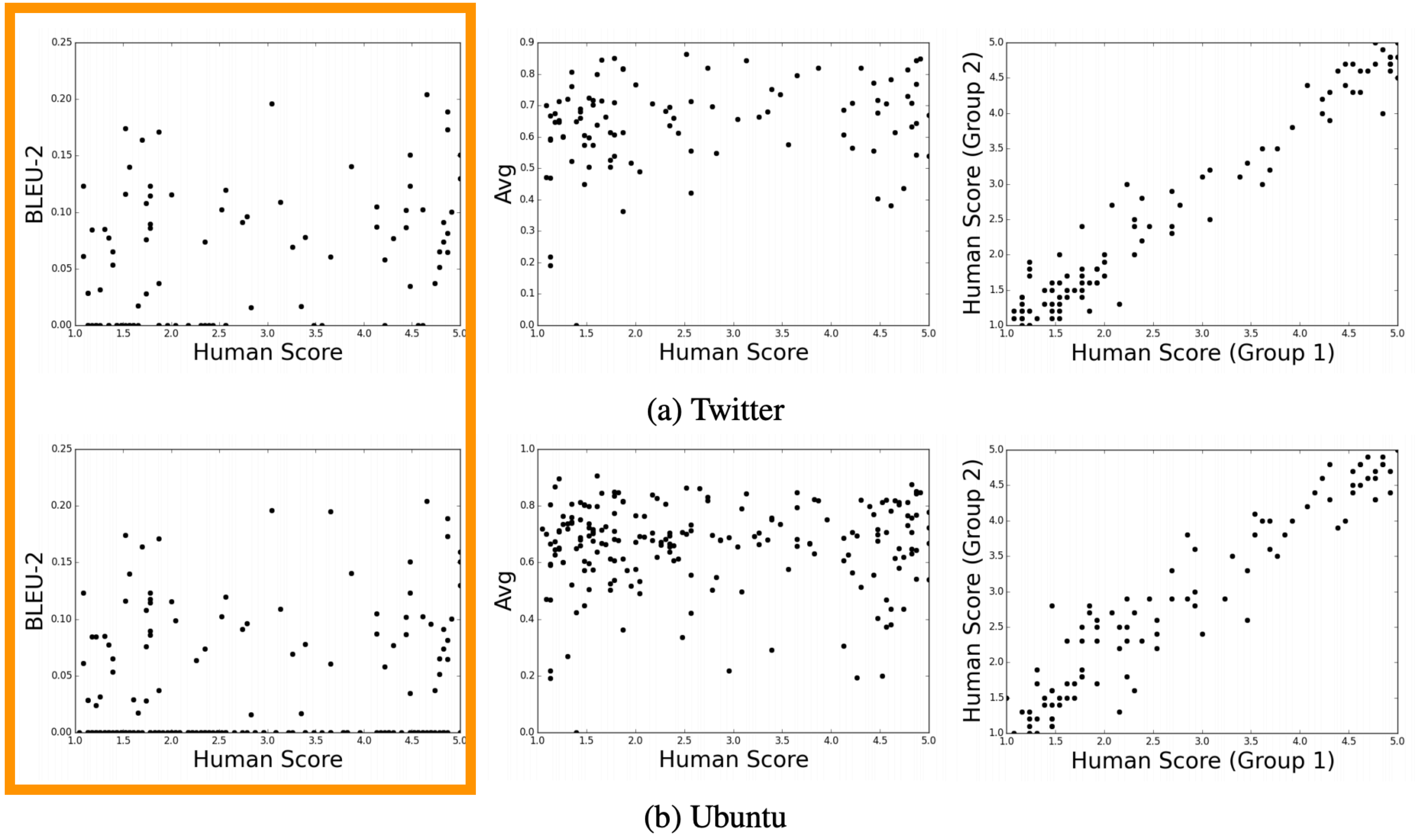
Model based metrics:

- Embedding similarity: Embedding average, Word Mover Distance, BERTSCORE, etc.
- Metric predictor: BLEURT

Issues with Automatic Evaluation

Automatic Evaluation:
Word overlap metrics are bad for dialogue

No correlation between **human judgement** and **BLEU**



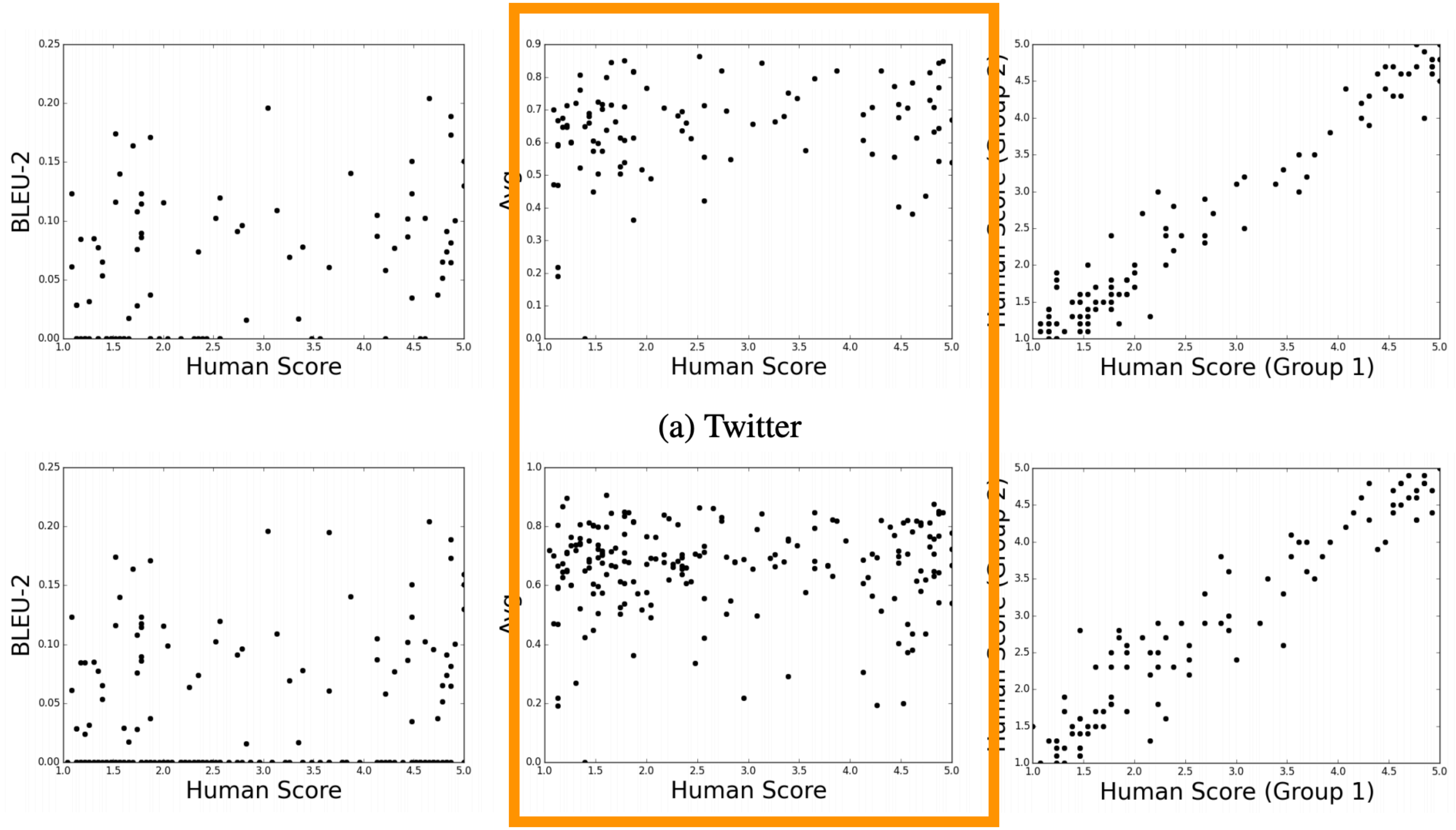
BLEU **Embedding Average** **Human**

[How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation, Liu et al 2017, https://arxiv.org/pdf/1603.08023.pdf]

Issues with Automatic Evaluation

Automatic Evaluation: Embedding metrics are also poor for dialogue

No correlation between **human judgement** and **embedding average**



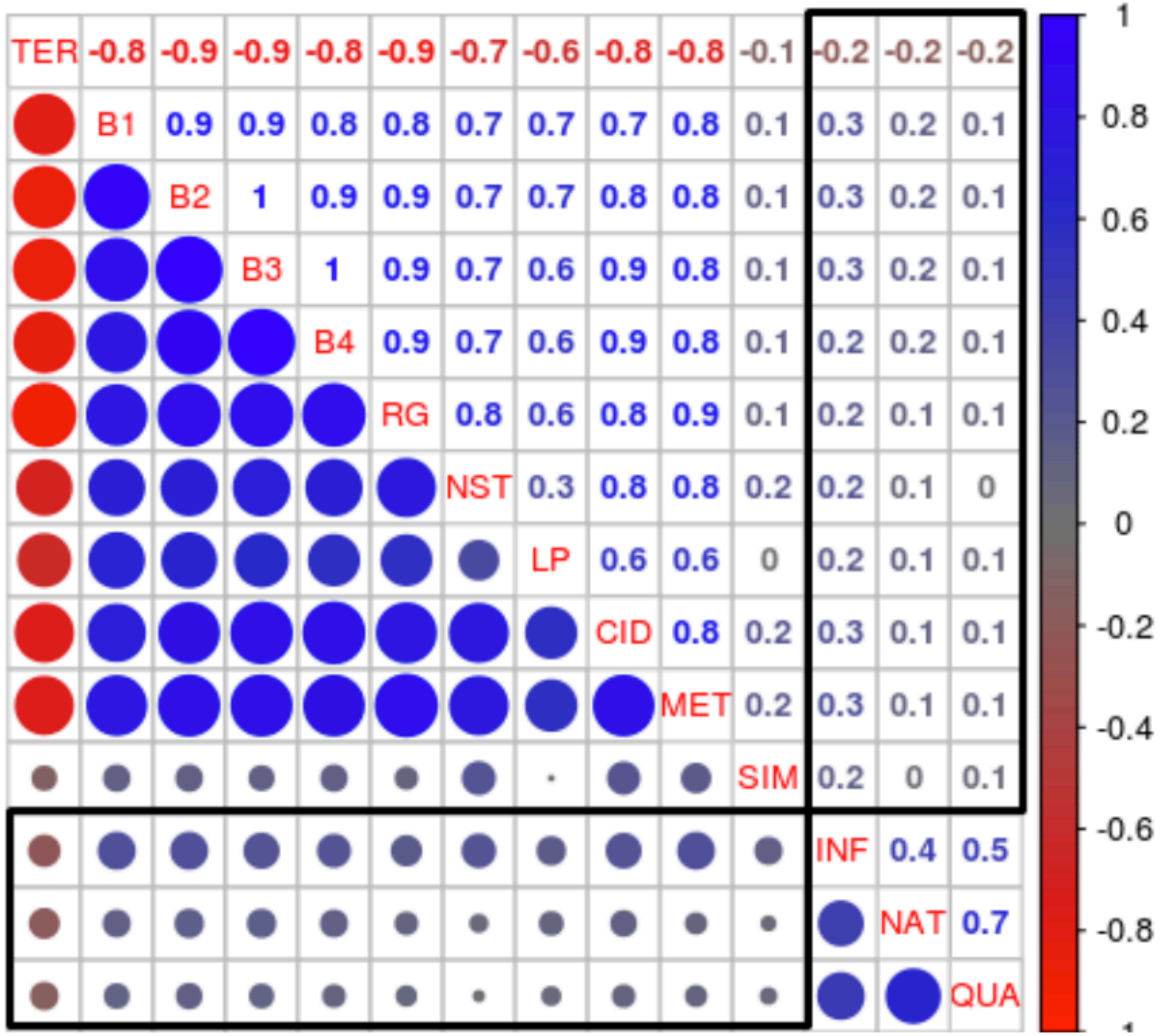
(b) Ubuntu

BLEU **Embedding Average** **Human**

[How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation, Liu et al 2017, https://arxiv.org/pdf/1603.08023.pdf]

Issues with Automatic Evaluation

Word Based Metrics



Spearman correlations of word based metrics and human ratings

Word Overlap Metrics

- highly correlated with each other
- Not so correlated with human ratings

Human Ratings

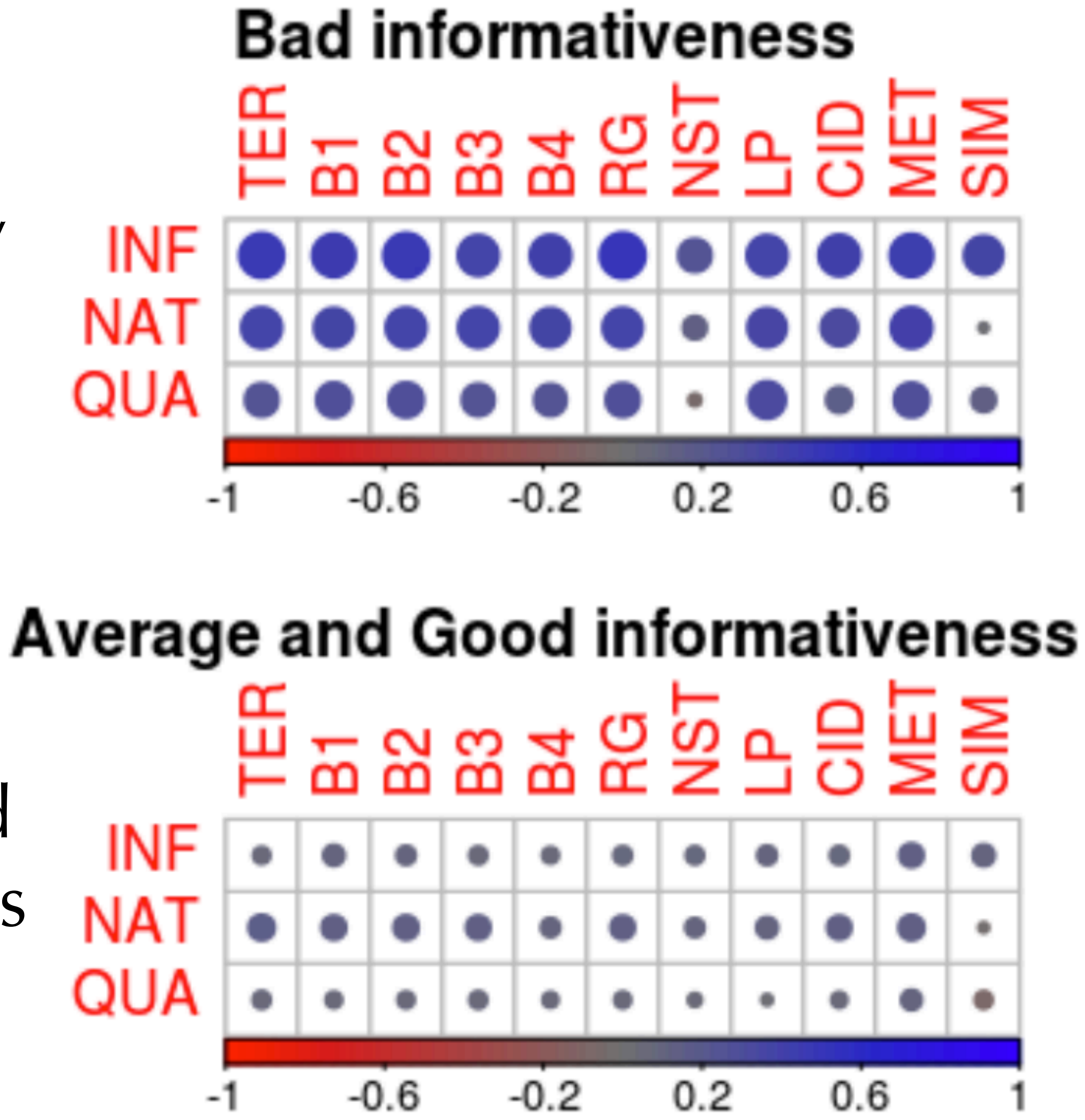
- Informativeness
- Naturalness
- Quality

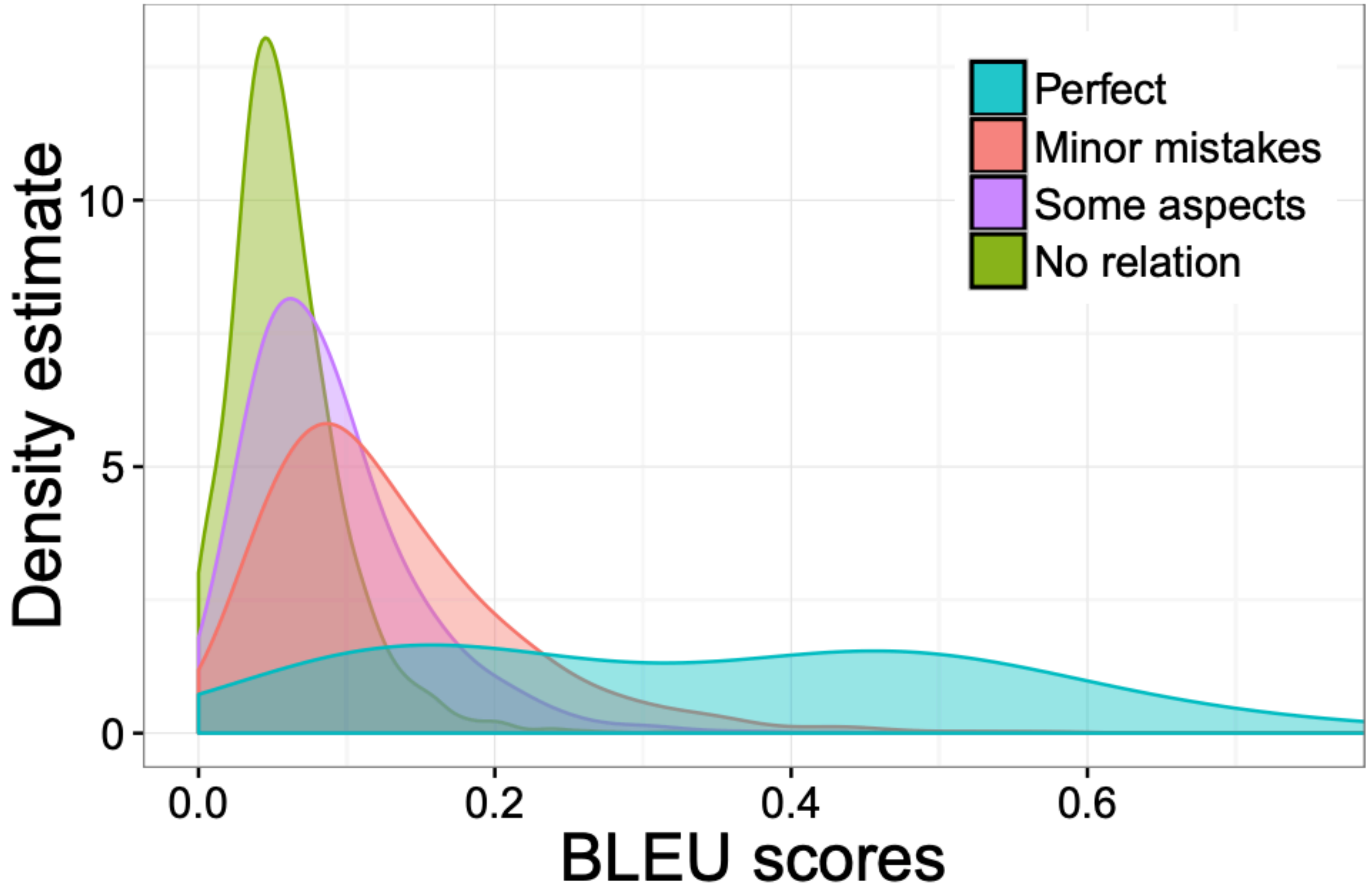
[Why We Need New Evaluation Metrics for NLG, Novikova et al 2017, https://arxiv.org/pdf/1707.06875.pdf]

Issues with automatic Evaluation

High correlation with human judgement for low quality generations

Poor correlation with human judgement for mid to high quality generations





[Why We Need New Evaluation Metrics for NLG, Novikova et al 2017, https://arxiv.org/pdf/1707.06875.pdf]

Re-evaluating Automatic Metrics for Image Captioning [Kilickaya et al, EACL 2017]

Human evaluation

What kind of human evaluation can be done?

- Can get ratings from chat **participants** or external **observers**.
- Can ask humans to rate various aspects of the chat (**likert scale**) or to compare two chats / responses (**AB testing**)
- Dimensions: fluency, coherence / consistency, factuality and correctness, commonsense, style / formality, grammaticality, typicality, redundancy

Issues with human evaluation

- slow, expensive
- not repeatable (subjective/inconsistent)
- difficult to form well-targeted questions that are not open to misinterpretation

When developing new automatic metrics, human evaluation is used as gold

- New automated metrics must correlate well with human evaluation.

Evaluation takeaways

- **Content overlap metrics** provide a good starting point for evaluating the quality of generated text, but they're not good enough on their own.
- **Model-based metrics** can be more correlated with human judgment, but metric may not be not interpretable
- **Human judgments** are critical
  - But humans are inconsistent and judgments are expensive
- If you are developing a NLG system, you should
  - Look at your model generations. Don't just rely on numbers!
  - Publicly release large samples of the output of systems that you create!

Slide adapted from Stanford CS224N (Xiang Lisa Li, Antoine Bosselut, Chris Manning)

Large language models for dialogue

ChatGPT can do cool stuff

juan
@juanbuis

having a particularly bizarre morning thanks to chatgpt

explain quantum theory to a child in the style of snoop dogg

Yo, little homie, let me break it down for you
Quantum theory, it's a crazy thing, it's true
See, the world is made up of tiny, tiny particles
That are always moving, and changing, and actin' wild

These particles, they don't always behave like we think
Sometimes they're waves, and sometimes they're points, it's a trip
They can be in two places at once, and they can be in two states
It's a mind-blowing concept, but that's the quantum fate

So don't be scared, little homie, it's all good
Quantum theory, it's just the way the world works, it's understood
It's a wild ride, but it's a cool one too
Quantum theory, it's the way the world does what it do.

ALT

5:33 AM · Dec 1, 2022

https://twitter.com/juanbuis/status/1598263964093251589?lang=en

ChatGPT

https://openai.com/blog/chatgpt/

Overall, very impressive

Question:

Recent models such as BERT (encoder-only) are applied mostly on classification tasks, whereas models like GPT1(decoder-only) are applied on generative modeling of text (including machine translation). Explain the justification of using these two different configurations of the transformer for the two tasks.

GPT:

The reason for using encoder-only models like BERT for classification tasks is that these tasks typically involve taking in an input sequence and outputting a single label or class for that sequence. In this case, the encoder self-attention blocks in BERT are able to capture the important relationships between the input elements and create a condensed representation of the input sequence, which is then passed to a classifier to make a prediction.

On the other hand, generative modeling tasks like machine translation require generating a sequence as output, which is why decoder-only models like GPT-1 are used. In these models, the decoder self-attention blocks are able to "pay attention" to the different elements of the encoded input sequence and previous decoder states in order to generate the next element in the output sequence, allowing the model to produce fluent and coherent text.

(from student on https://coursys.sfu.ca/2022fa-cmpt-413-x1/forum/386)

ChatGPT

- ChatGPT: released Nov 30th 2022, 1M users in 5 days, 100M users in 2 months

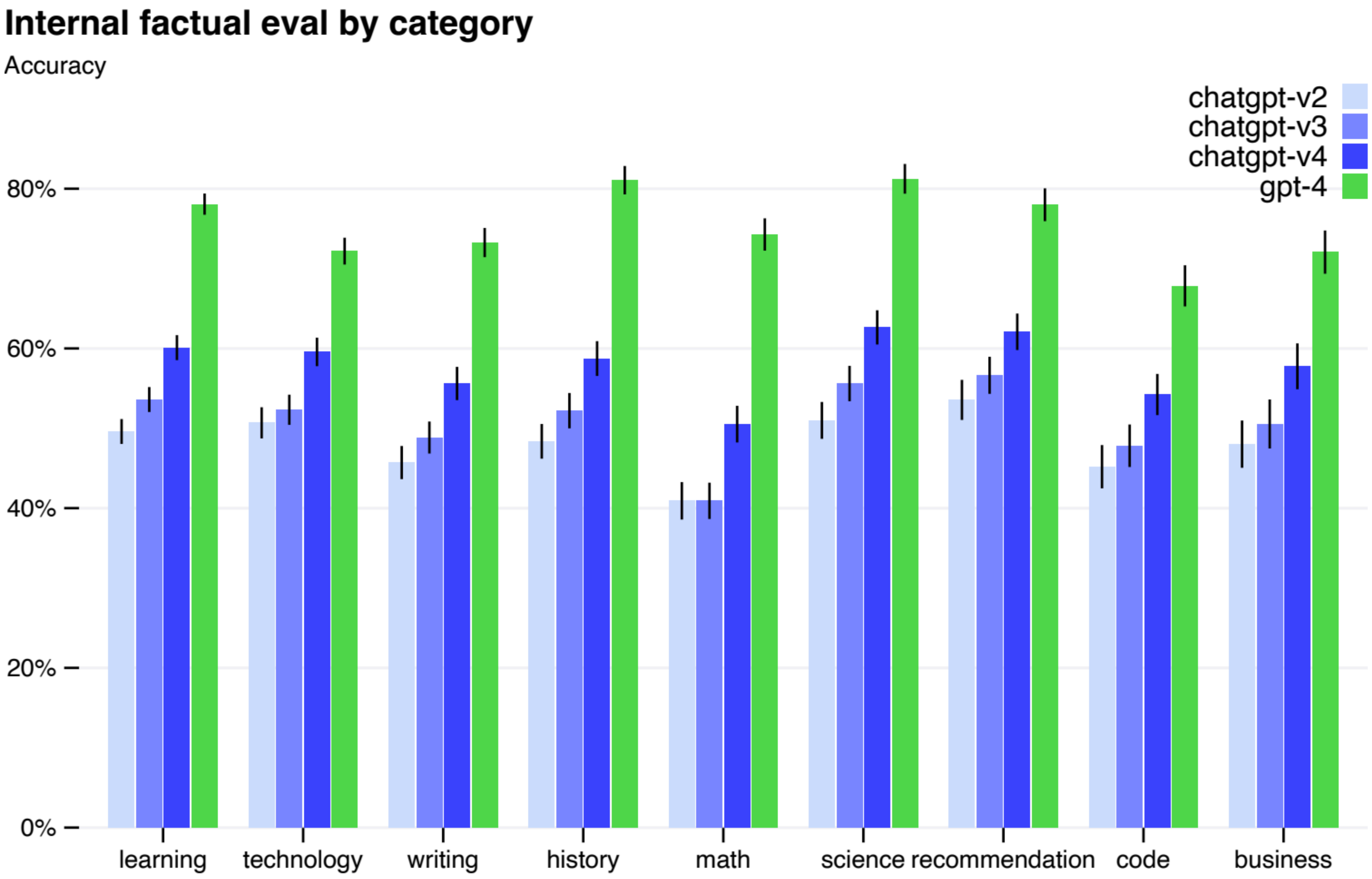
- Large language model (GPT-3.5)
  - GPT-3 (2020) - 175B parameter model
  - GPT-3.5 (late 2022) - trained on a blend of text and code
  - GPT-4 (March 2023) - multimodal

- Supervised fine-tuning on human conversations
  - Data where human will pretend to be user or AI assistant

- Reinforcement learning with human feedback
  - Humans rank what response is best
  - Aim: reduce harmful / deceitful responses

GPT-4

- Growing performance for ChatGPT versions

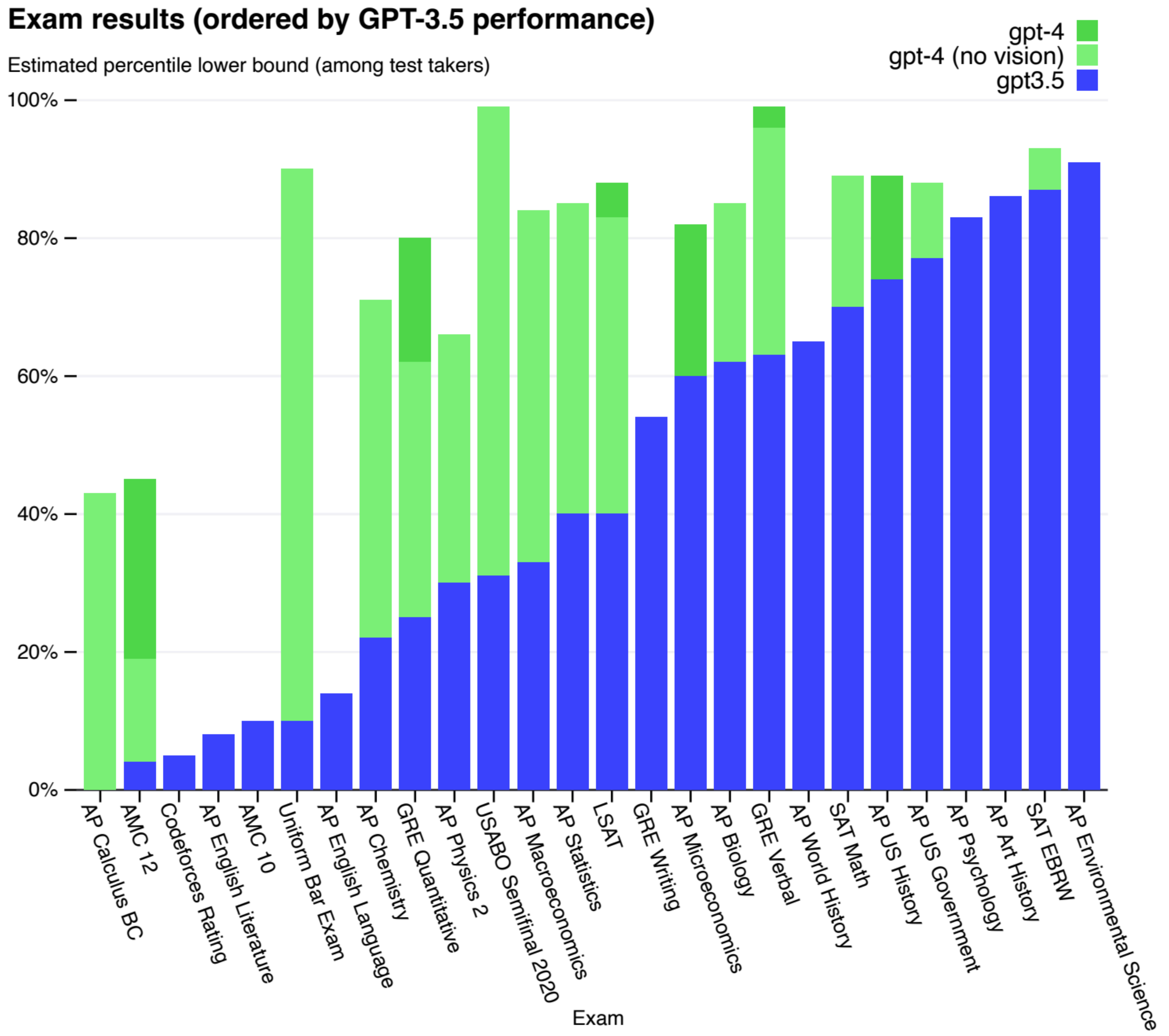


https://openai.com/research/gpt-4

GPT-4

- GPT-4 passing standardized tests

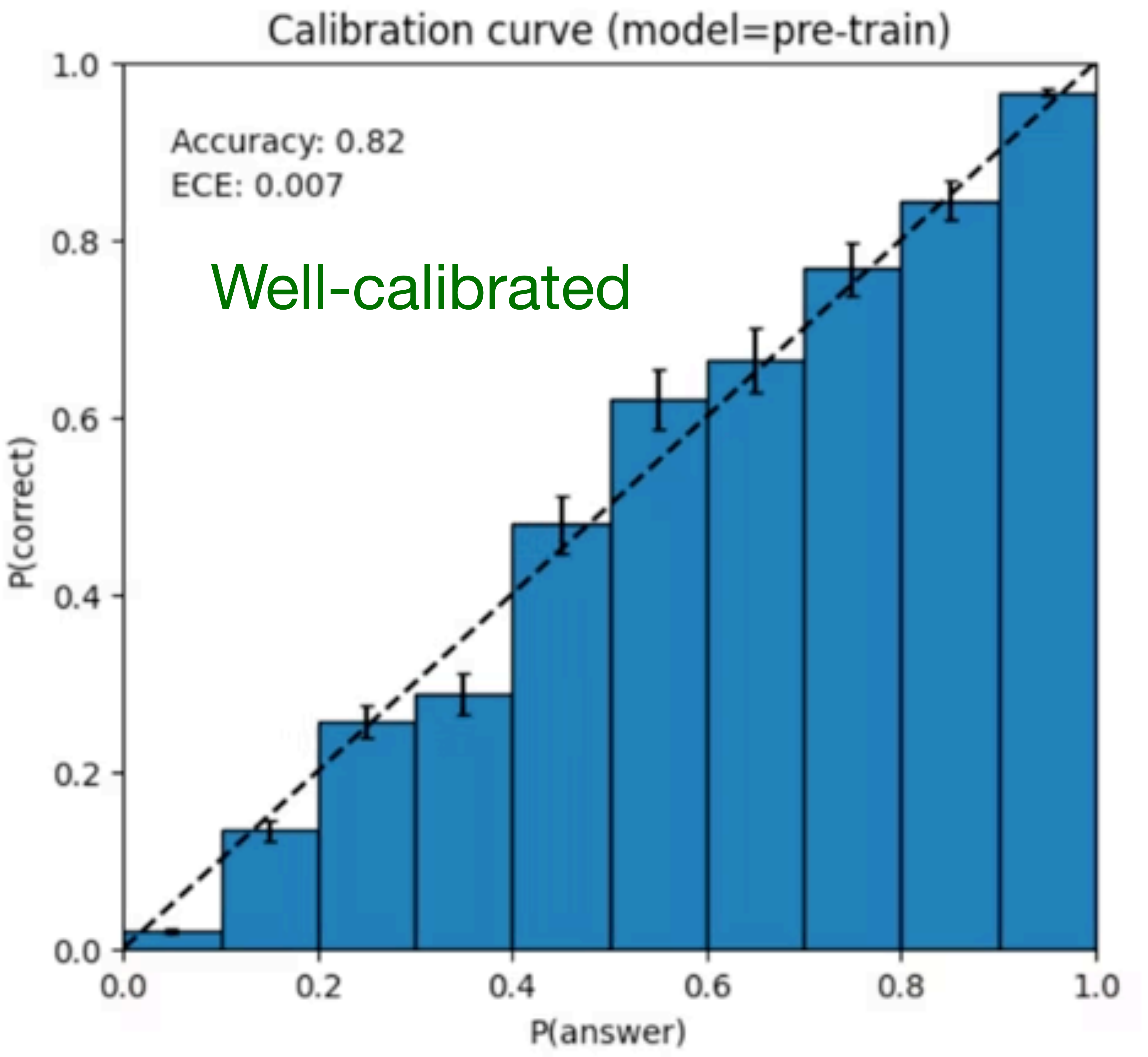
- Bar exam:
  - GPT-3.5 score in bottom 10%
  - GPT-4 score in top 10%

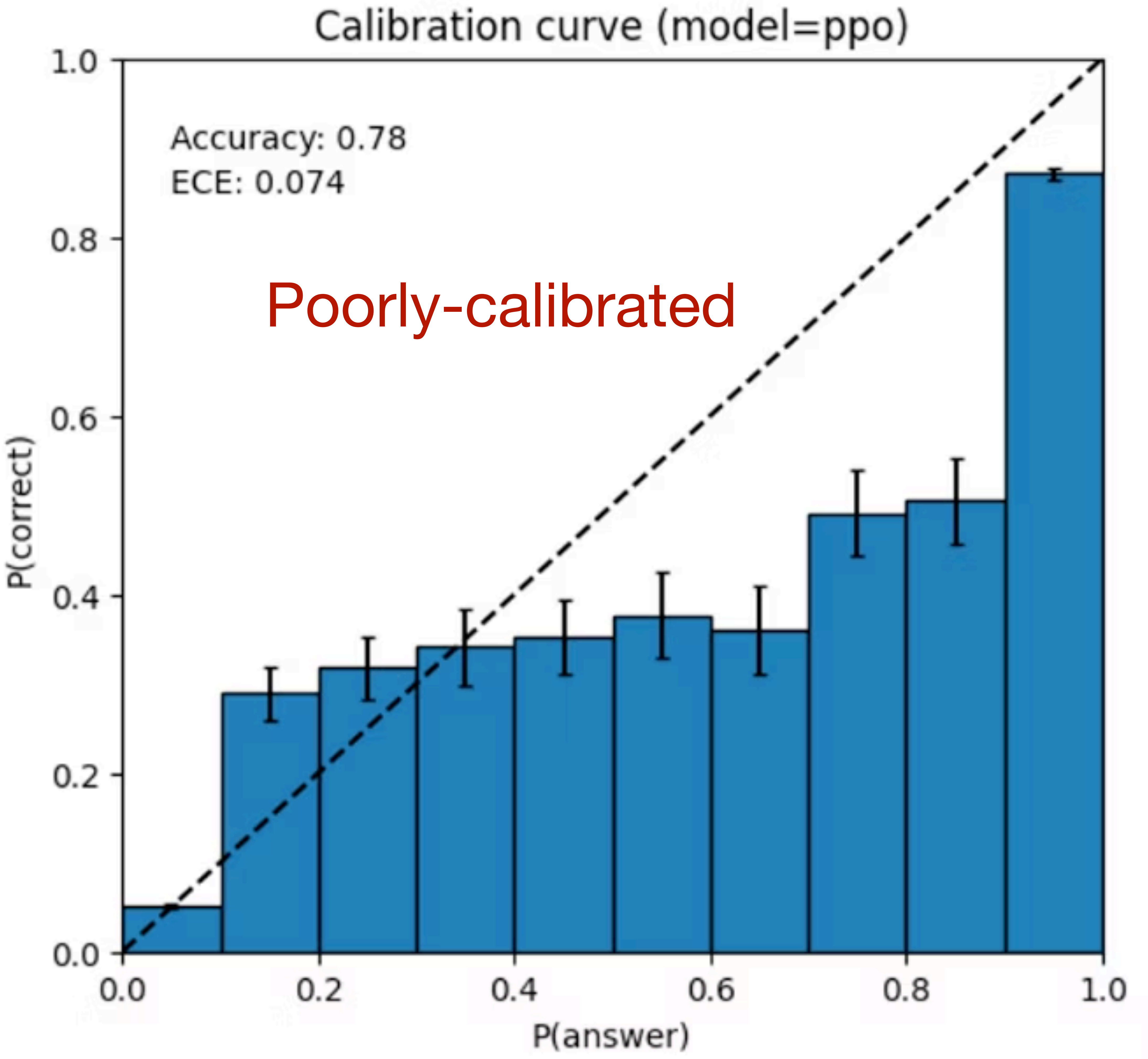


https://openai.com/research/gpt-4

Pre-trained GPT-4 is well-calibrated

Calibrated: predicted confidence matches probability of being correct





https://openai.com/research/gpt-4

ChatGPT

- ChatGPT: released Nov 30th 2022, 1M users in 5 days, 100M users in 2 months

- Large language model (GPT-3.5)
  - GPT-3 (2020) - 175B parameter model
  - GPT-3.5 (late 2022) - trained on a blend of text and code
  - GPT-4 (March 2023) - multimodal

- Supervised fine-tuning on human conversations
  - Data where human will pretend to be user or AI assistant

- Reinforcement learning with human feedback
  - Humans rank what response is best
  - Aim: reduce harmful / deceitful responses

InstructGPT

Collect data from humans

Example user prompts

open qa

Q: Who is Batman?
A: Batman is a fictional comic book character.

Q: What is torsalplexity?
A: ?

Q: What is Devz9?
A: ?

Q: Who is George Lucas?
A: George Lucas is American film director and producer famous for creating Star Wars.

Q: What is the capital of California?
A:

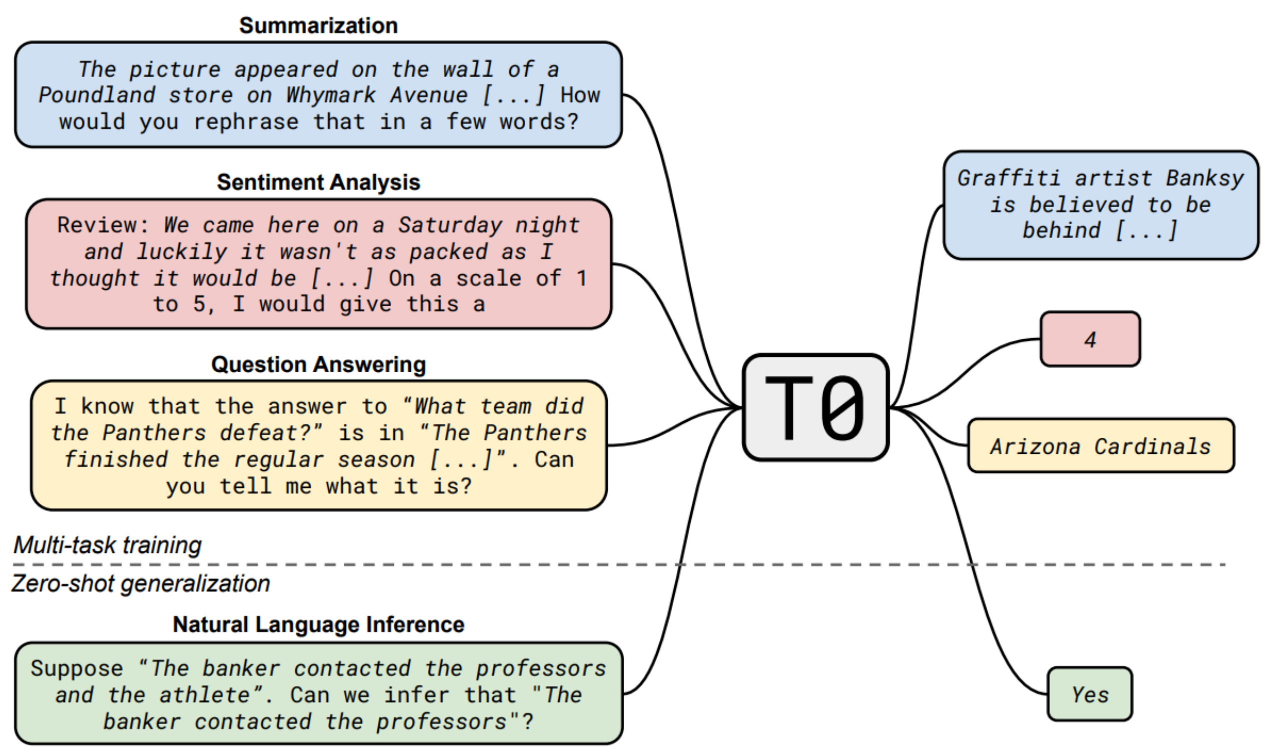
chat

The following is a conversation with an AI assistant. The assistant is helpful, creative, clever, and very friendly.

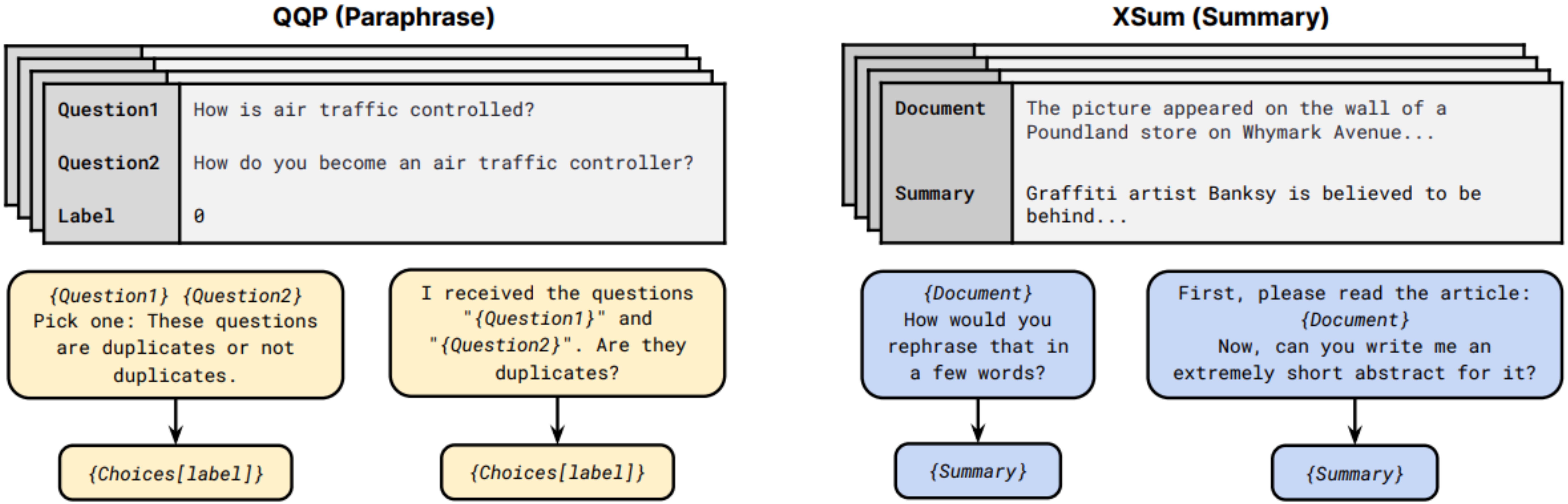
Human: Hello, who are you?
AI: I am an AI created by OpenAI. How can I help you today?
Human: I'm feeling kind of down today.
AI:

Training language models to follow instructions with human feedback, Ouyang et. al. OpenAI 2022

Instruction tuning

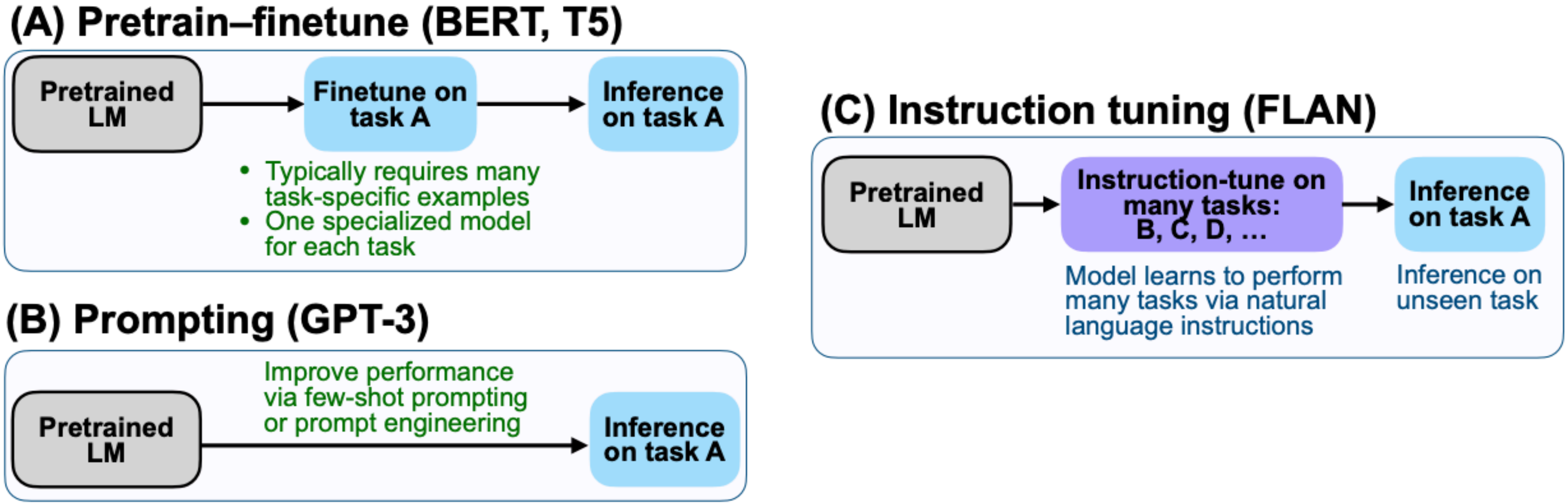


- Use templates to make them into instruction based dataset
- Text based format makes it natural for humans



Multitask prompted training enables zero-shot task generalization, Sahn et al. Google, ICLR 2022

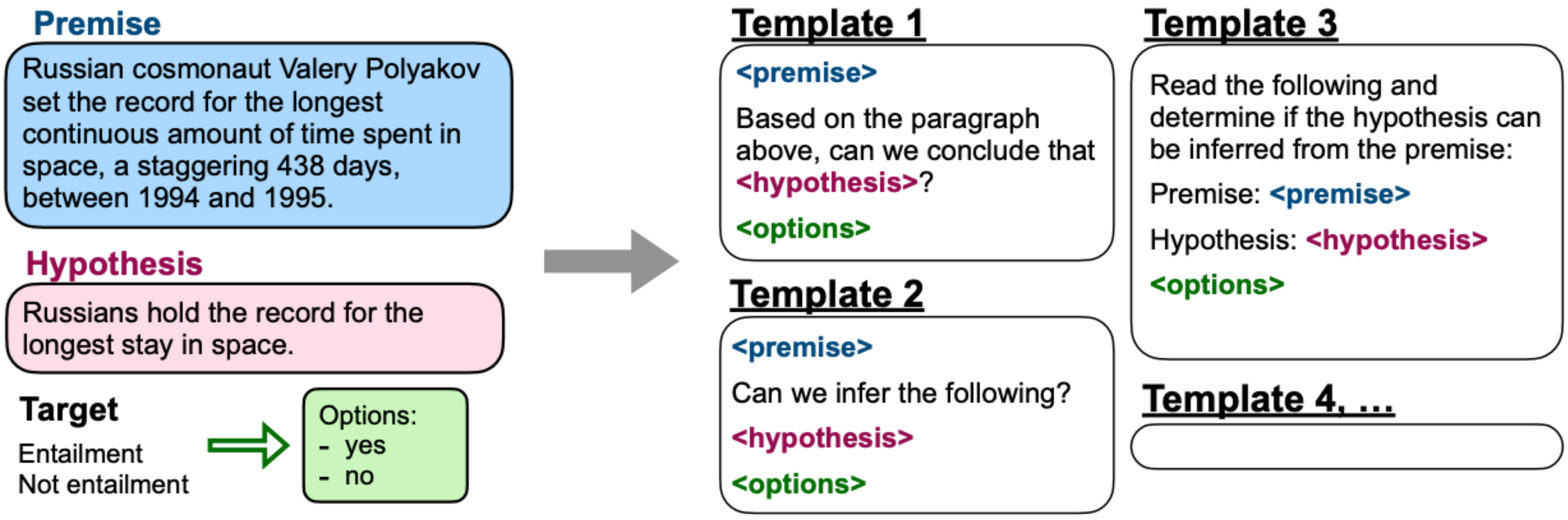
Instruction tuning



Finetuned language models are zero-shot learners, Wei et al. Google, ICLR 2022

Instruction tuning

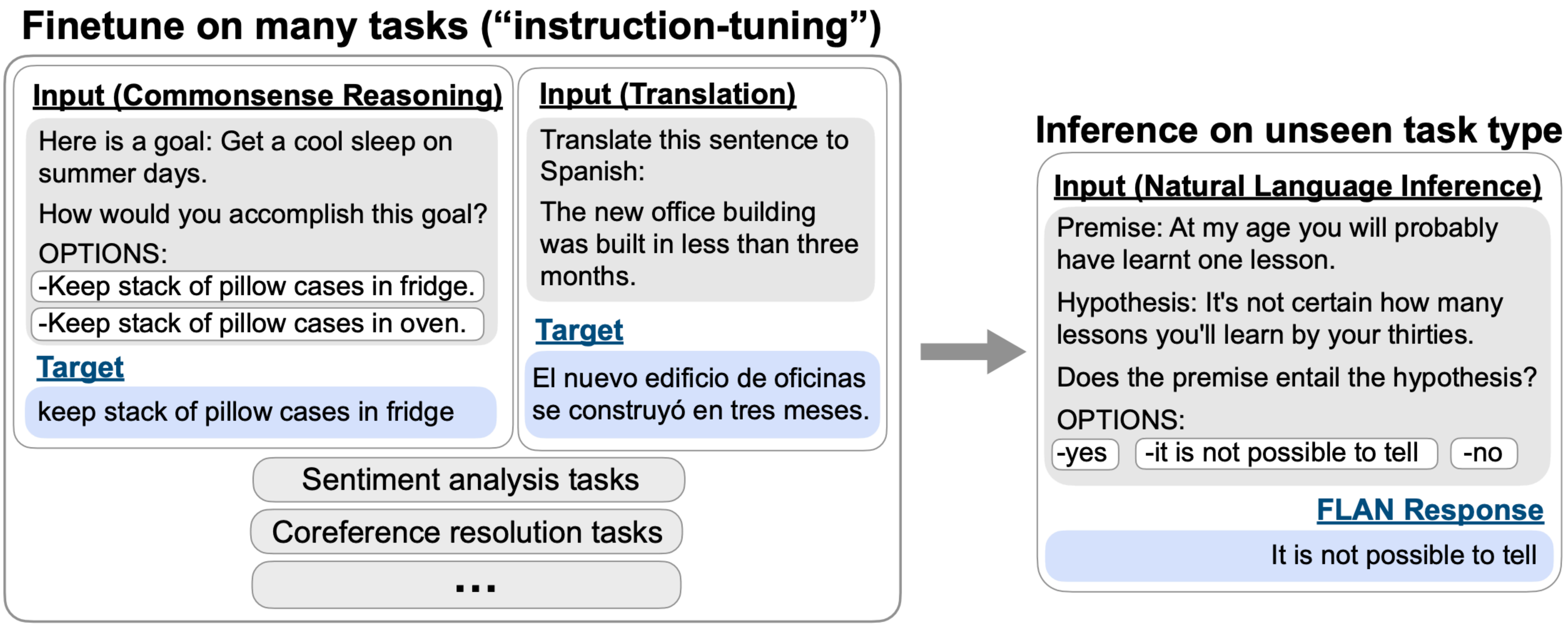
- Use templates to make them into instruction based dataset
- Text based format makes it natural for humans



Finetuned language models are zero-shot learners, Wei et al. Google, ICLR 2022

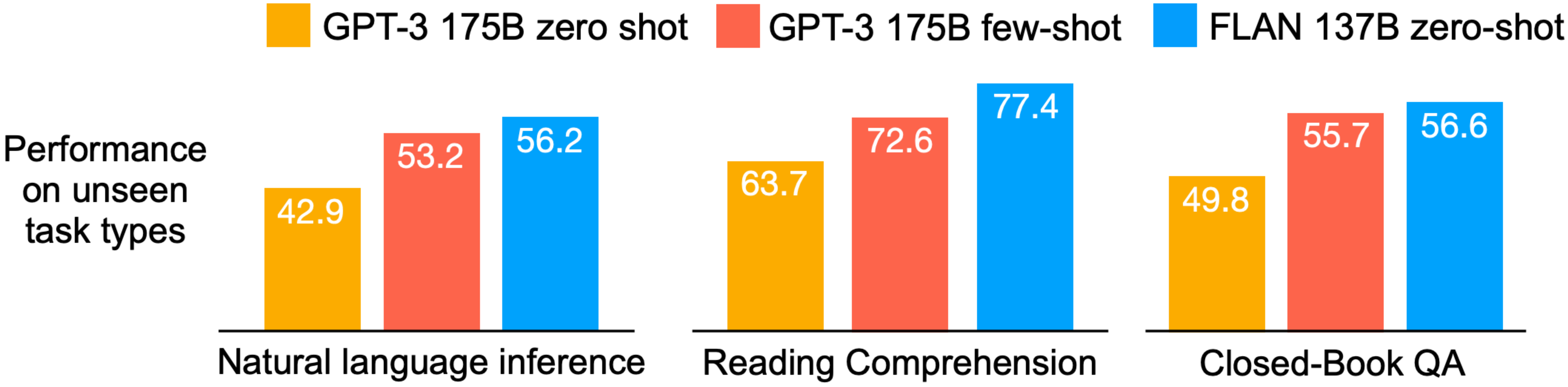
Instruction tuning

- Can be used on an unseen task type



Instruction tuning

- Can be used on an unseen task type



Finetuned language models are zero-shot learners, Wei et al. Google, ICLR 2022

ChatGPT

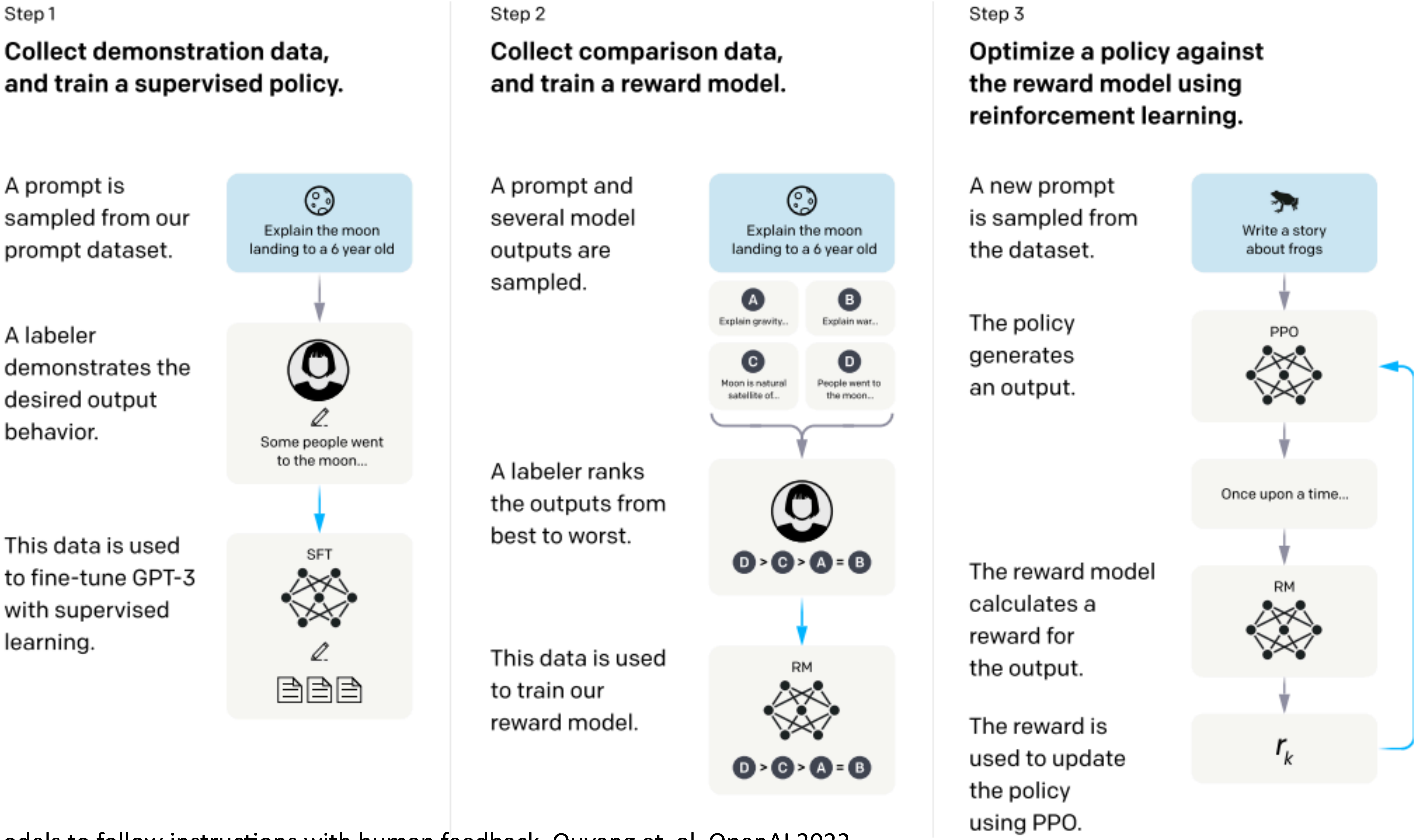
- ChatGPT: released Nov 30th 2022, 1M users in 5 days, 100M users in 2 months

- Large language model (GPT-3.5)
  - GPT-3 (2020) - 175B parameter model
  - GPT-3.5 (late 2022) - trained on a blend of text and code
  - GPT-4 (March 2023) - multimodal

- Supervised fine-tuning on human conversations
  - Data where human will pretend to be user or AI assistant

- Reinforcement learning with human feedback
  - Humans rank what response is best
  - Aim: reduce harmful / deceitful responses

InstructGPT



Training language models to follow instructions with human feedback, Ouyang et. al. OpenAI 2022

InstructGPT

Collect human judgement of which responses are better

- Labelers rank K (K=4 to 9) responses for each prompt
- Gives $\binom{K}{2}$ comparisons per prompt

Train reward model (RM) r_θ to over human data *D*:

$$\operatorname{loss}(\theta) = -\frac{1}{\binom{K}{2}} E_{(x, y_w, y_l) \sim D} \left[\log\left(\sigma\left(r_\theta\left(x, y_w\right) - r_\theta\left(x, y_l\right)\right)\right)\right]$$

Use reinforcement learning (RL) to train RL based policy for selecting words to generate

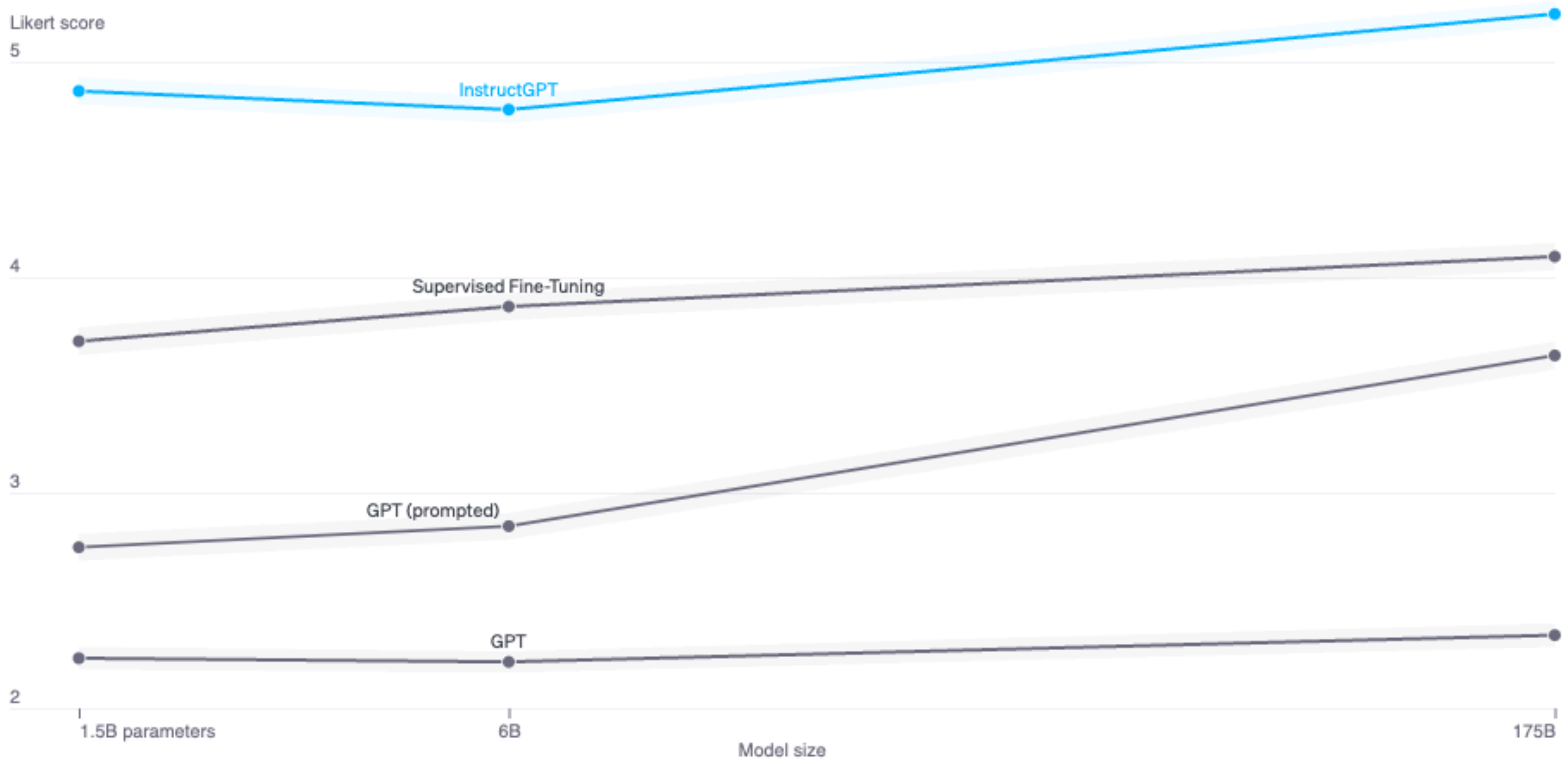
$$\begin{aligned}\operatorname{objective}(\phi) = & E_{(x,y) \sim D_{\pi_\phi^{\mathrm{RL}}}} \left[r_\theta(x, y) - \beta \log\left(\pi_\phi^{\mathrm{RL}}(y \mid x) / \pi^{\mathrm{SFT}}(y \mid x)\right)\right] + \\ & \gamma E_{x \sim D_{\text{pretrain}}} \left[\log(\pi_\phi^{\mathrm{RL}}(x))\right]\end{aligned}$$

Training language models to follow instructions with human feedback, Ouyang et. al. OpenAI 2022

InstructGPT

Aligns model behaviour to human preferences

Human judgement of model output



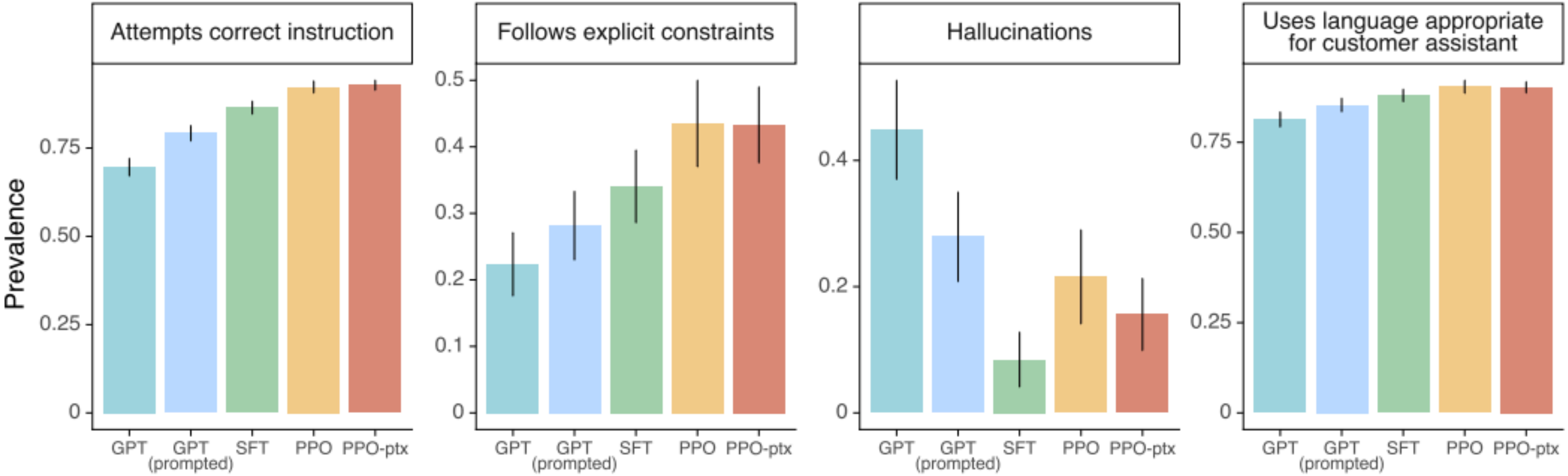
Training language models to follow instructions with human feedback, Ouyang et. al. OpenAI 2022

https://openai.com/research/instruction-following

InstructGPT

Aligns model behaviour to human preferences

Human judgement of model output



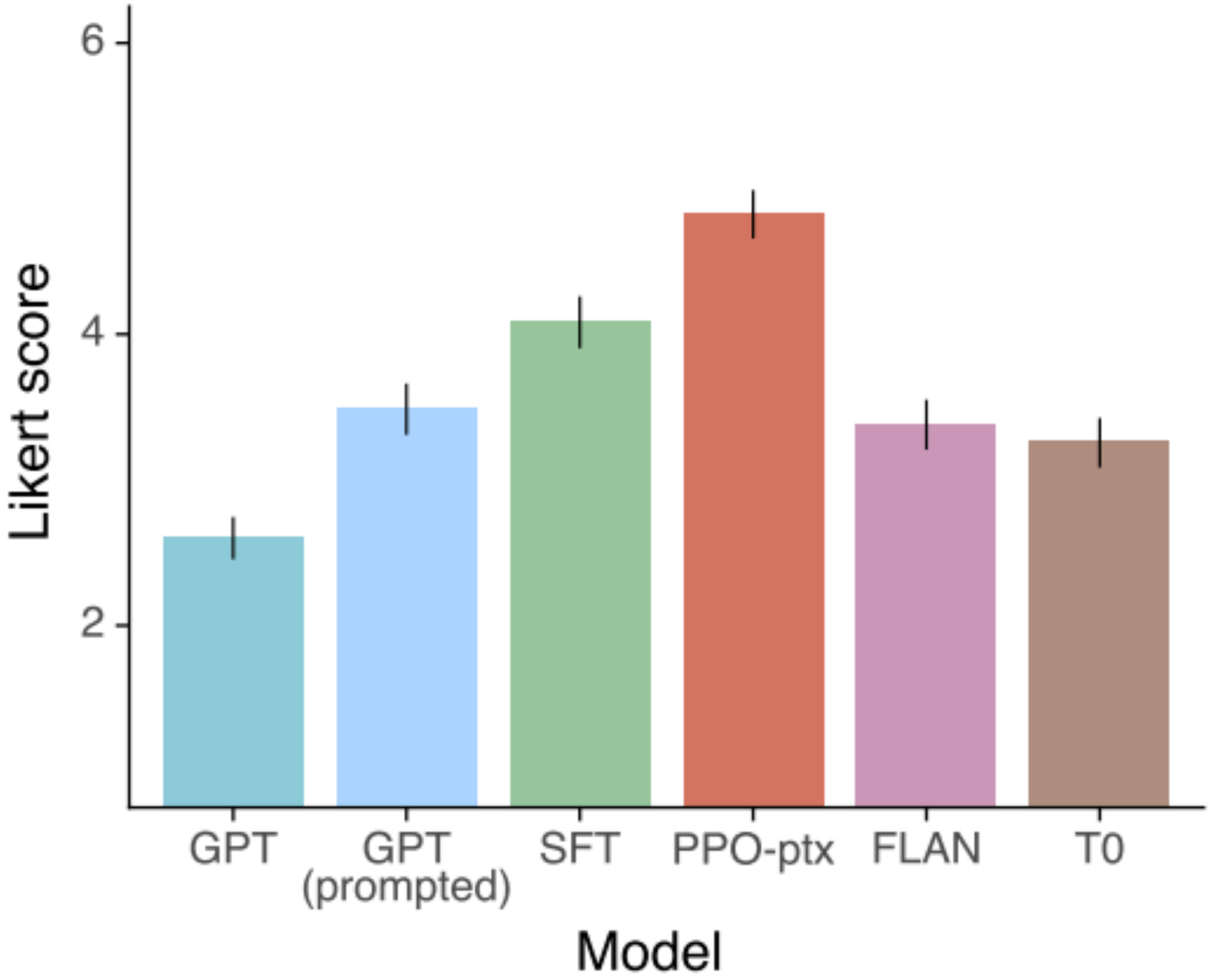
Training language models to follow instructions with human feedback, Ouyang et. al. OpenAI 2022

https://openai.com/research/instruction-following

InstructGPT

Aligns model behaviour to human preferences

Human judgement of model output

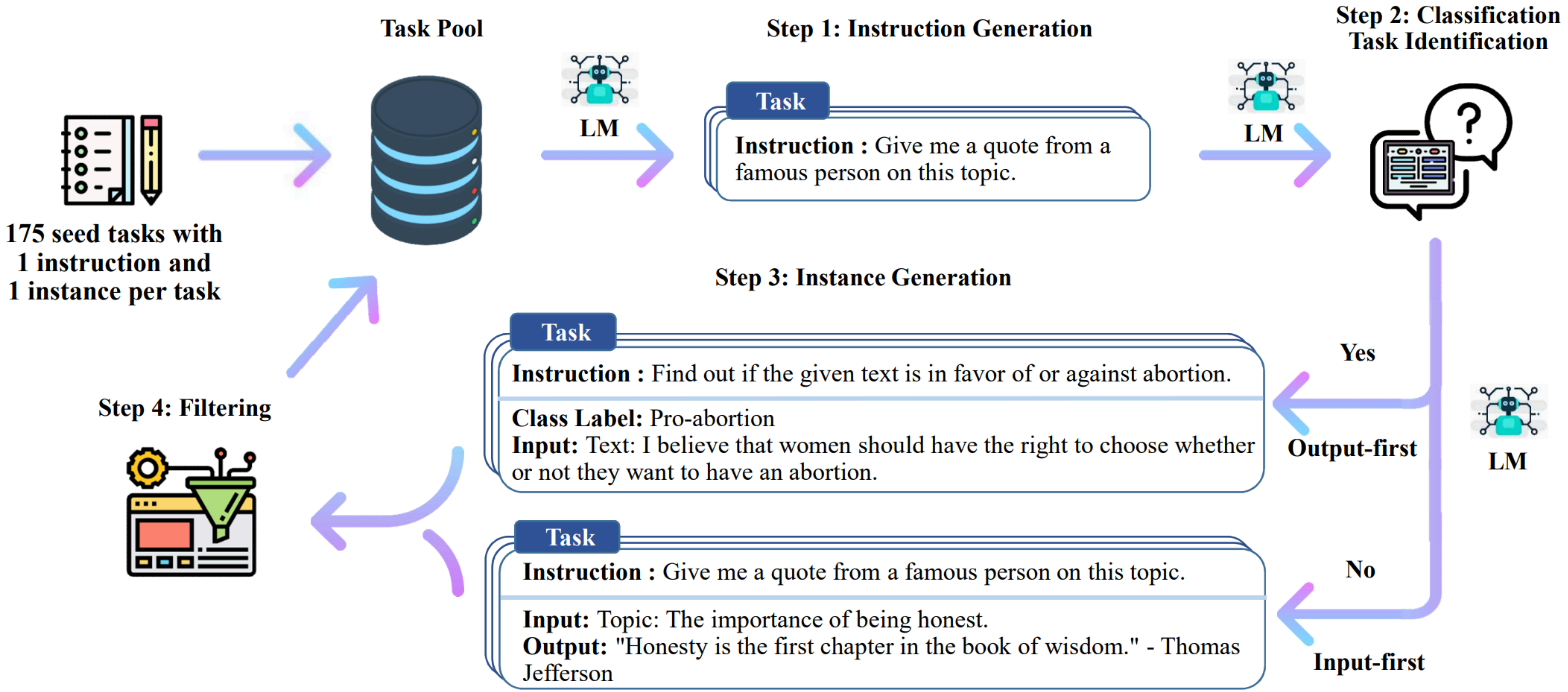


Training language models to follow instructions with human feedback, Ouyang et. al. OpenAI 2022

https://openai.com/research/instruction-following

Self-Instruct

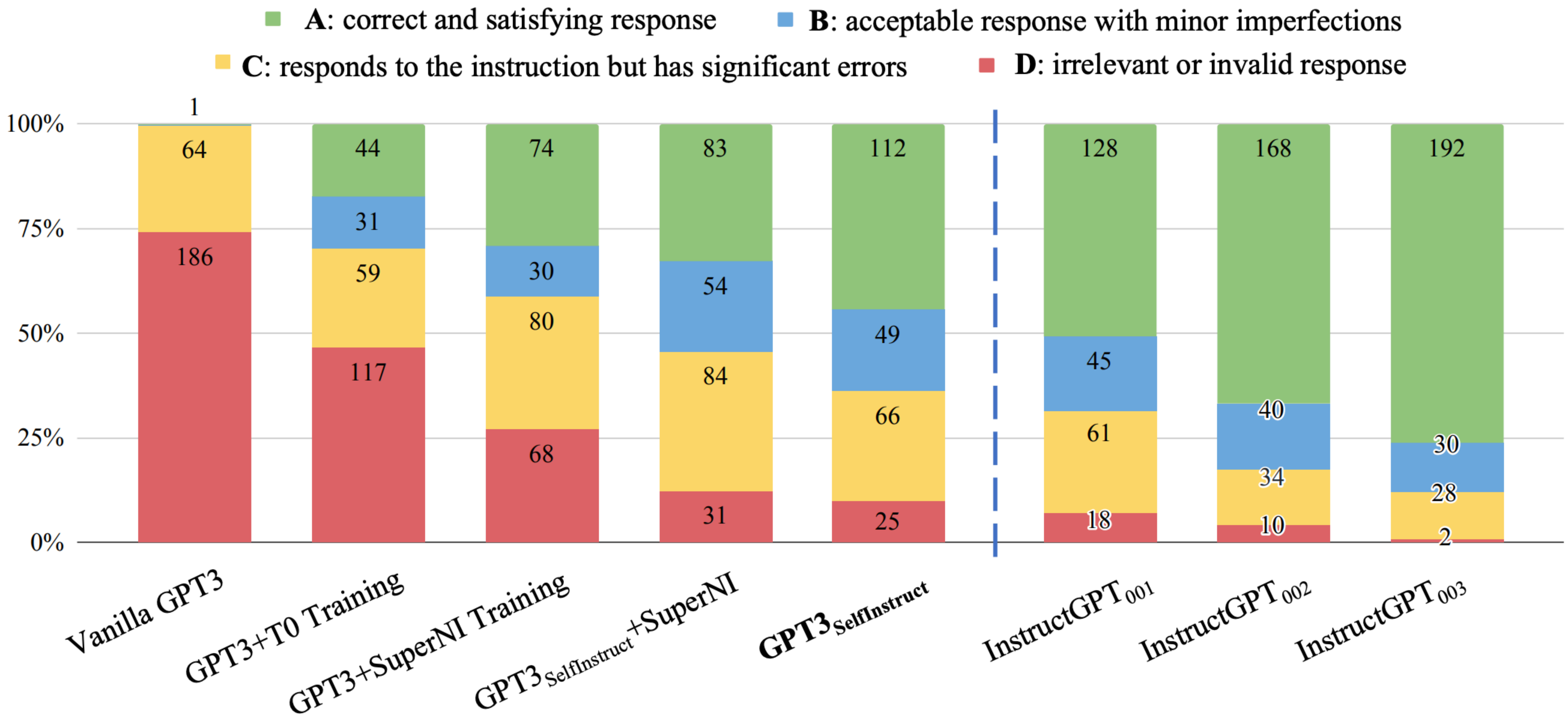
- Generate task instructions using LLMs to train/fine-tune LLMs!



SELF-INSTRUCT: Aligning Language Model with Self Generated Instructions, Wang et al. 2022

Self-Instruct

- Generate task instructions using LLMs to train/fine-tune LLMs!



SELF-INSTRUCT: Aligning Language Model with Self Generated Instructions, Wang et al. 2022

Self-Instruct

Text-davinci-003

175 Self-Instruct seed tasks

Modified Self-instruct Instruction Generation

Meta

LLaMA 7B

52K Instruction-following examples

Supervised Finetuning

Alpaca 7B

Example seed task

Instruction: Brainstorm a list of possible New Year's resolutions.

Output:
- Lose weight
- Exercise more
- Eat healthier

Example Generated task

Instruction: Brainstorm creative ideas for designing a conference room.

Output:
... incorporating flexible components, such as moveable walls and furniture ...