Evalution

Natural Language Processing — Lecture 9

Kenneth Enevoldsen | 2024

CENTER FOR HUMANITIES COMPUTING

Learning goals

- Be able to relate evaluation to existing knowledge on evaluation in machine learning
- The student should be able to choose the right evaluation method for a given question
- Have a reasonable overview of methods of evaluation within NLP including quantitative, qualitative and mixed approaches
  - Have an understanding of the limitations of evaluation methods
- Students should be able to examine the failure modes of a system

Quiz!

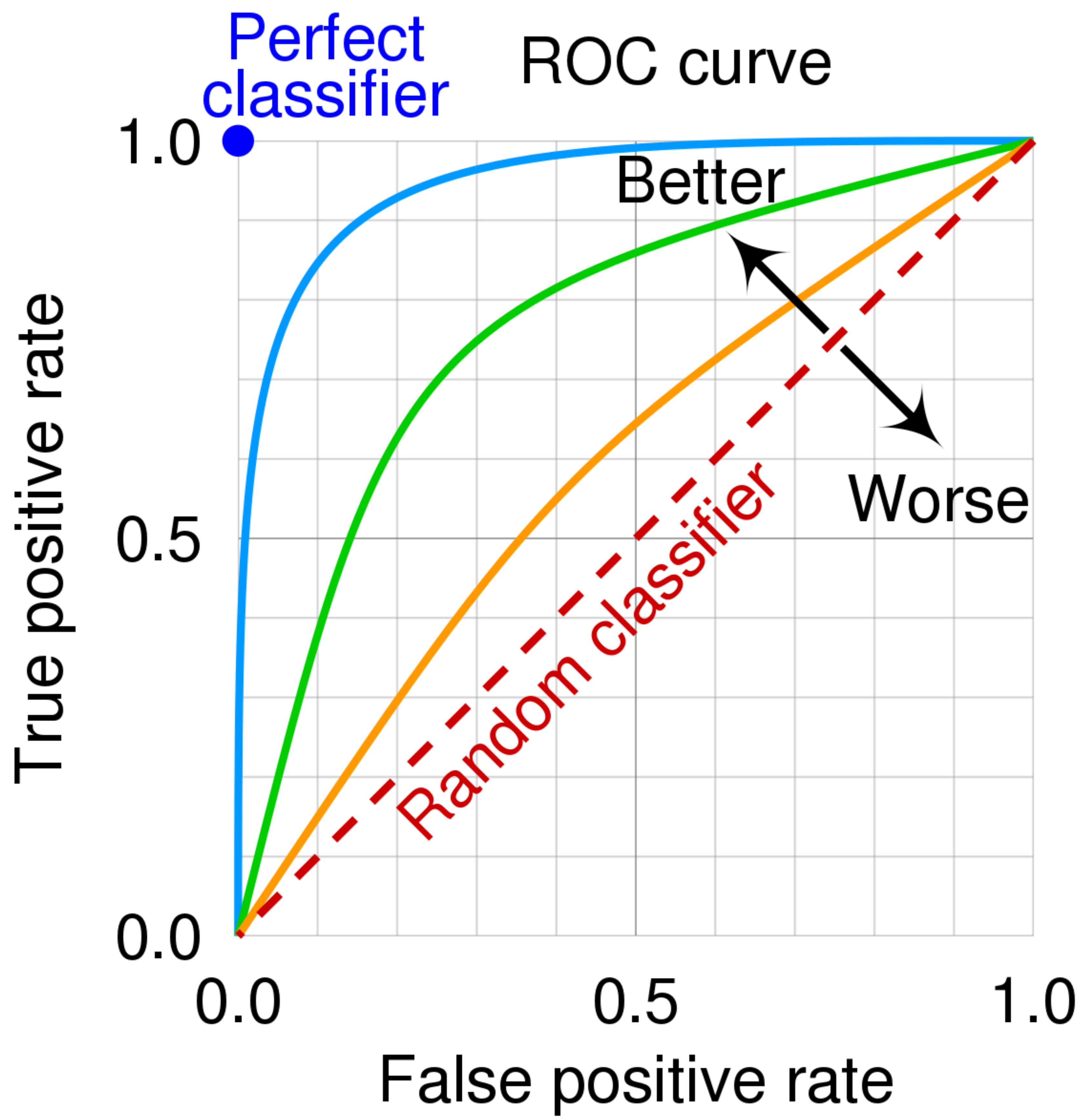
- https://www.menti.com/alhjpunv9ay4

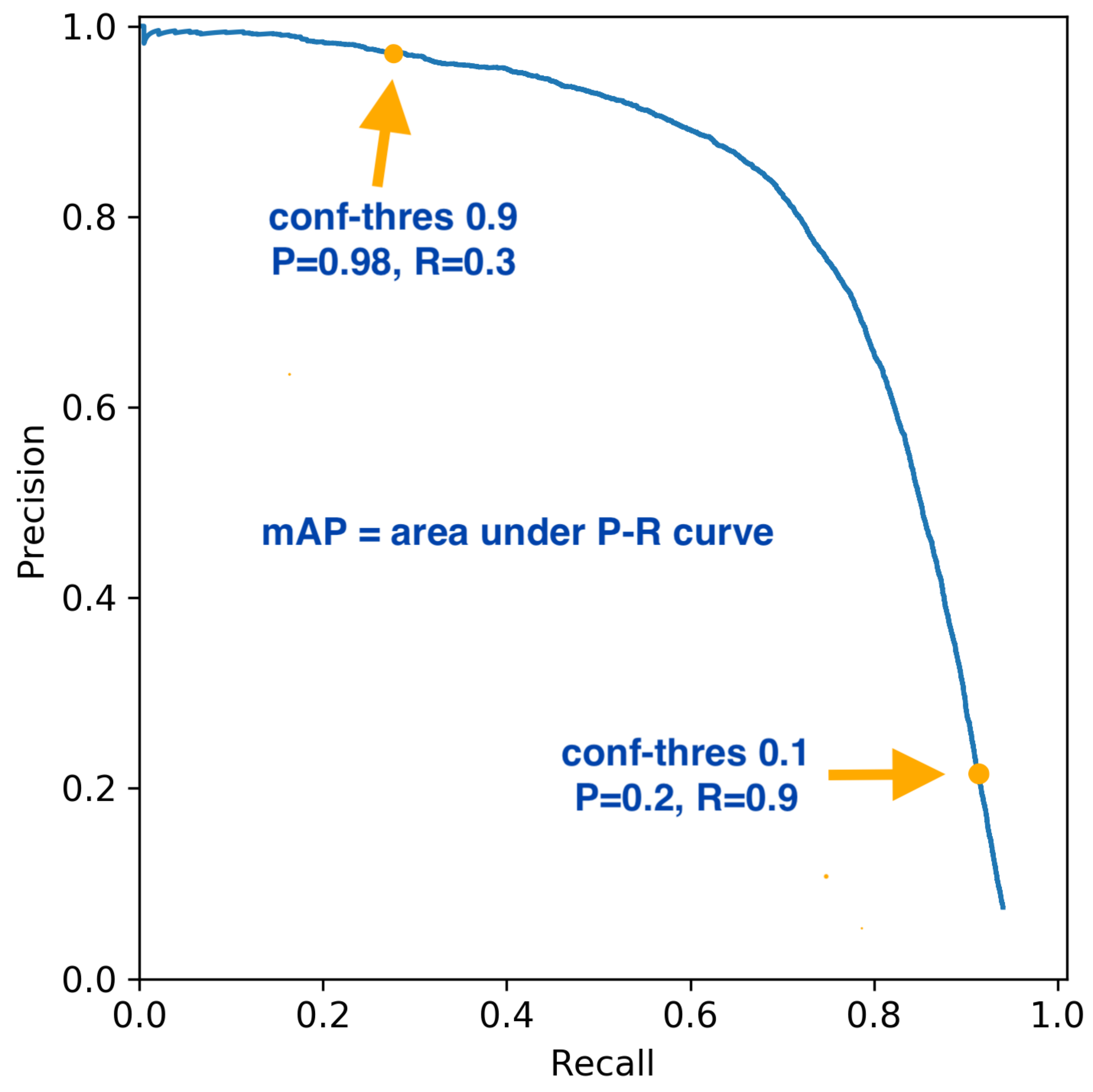
Recap: Machine learning Evaluations

- Accuracy = $\dfrac{TP + TN}{TP + FP + TN + FN}$
- Precision = $\dfrac{TP}{TP + TP}$
- Recall = $\dfrac{TP}{TP + FN}$
- F1 score = (harmonic) mean of precision and recall

		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

ROC AUC and the precision recall curve





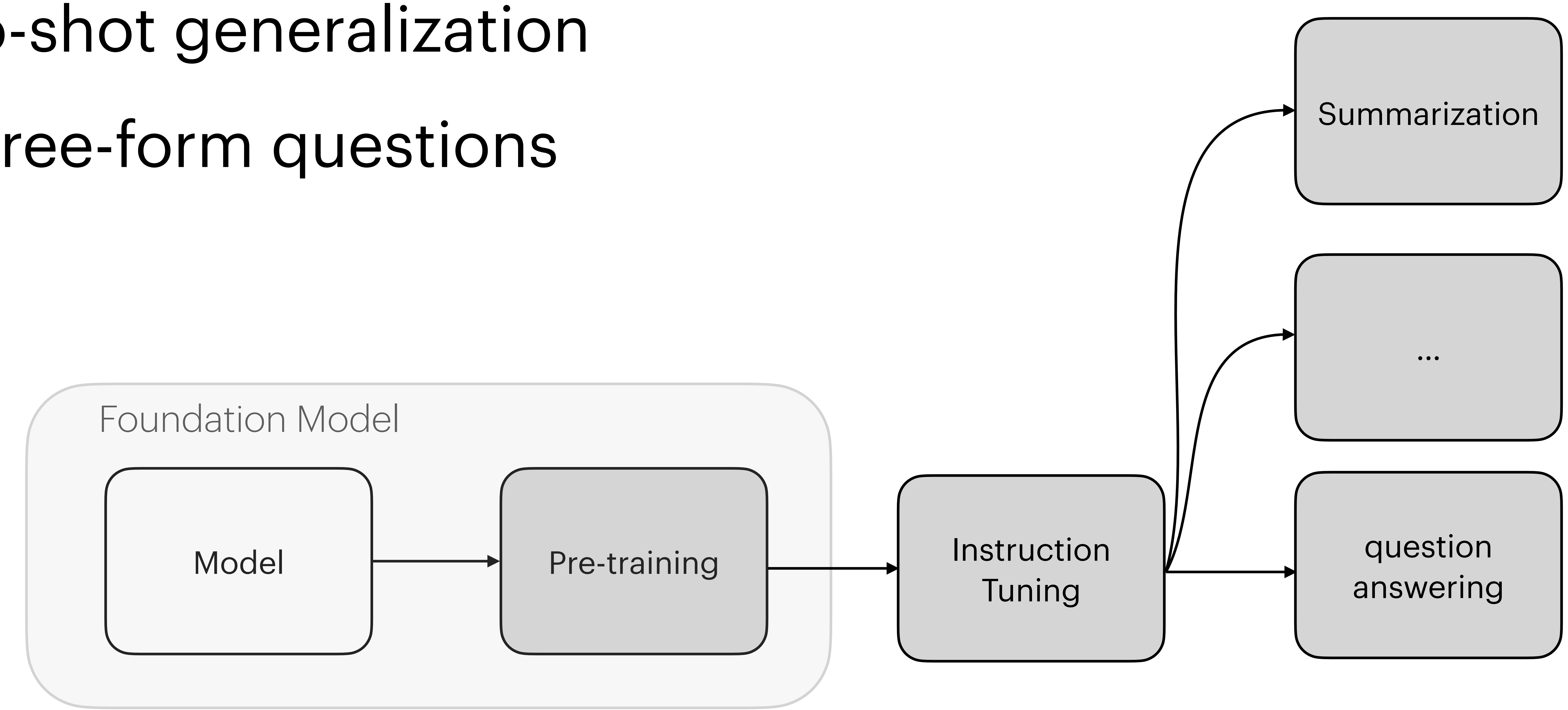
Different tasks calls for different measures

In fact, the Chinese NORP market has the three CARDINAL most influential names of the retail and tech space – Alibaba GPE , Baidu ORG , and Tencent PERSON (collectively touted as BAT ORG), and is betting big in the global AI GPE in retail industry space . The three CARDINAL giants which are claimed to have a cut-throat competition with the U.S. GPE (in terms of resources and capital) are positioning themselves to become the 'future AI PERSON platforms'. The trio is also expanding in other Asian NORP countries and investing heavily in the U.S. GPE based AI GPE startups to leverage the power of AI GPE . Backed by such powerful initiatives and presence of these conglomerates, the market in APAC AI is forecast to be the fastest-growing one CARDINAL , with an anticipated CAGR PERSON of 45% PERCENT over 2018 - 2024 DATE .

To further elaborate on the geographical trends, North America LOC has procured more than 50% PERCENT of the global share in 2017 DATE and has been leading the regional landscape of AI GPE in the retail market. The U.S. GPE has a significant credit in the regional trends with over 65% PERCENT of investments (including M&As, private equity, and venture capital) in artificial intelligence technology. Additionally, the region is a huge hub for startups in tandem with the presence of tech titans, such as Google ORG , IBM ORG , and Microsoft ORG .

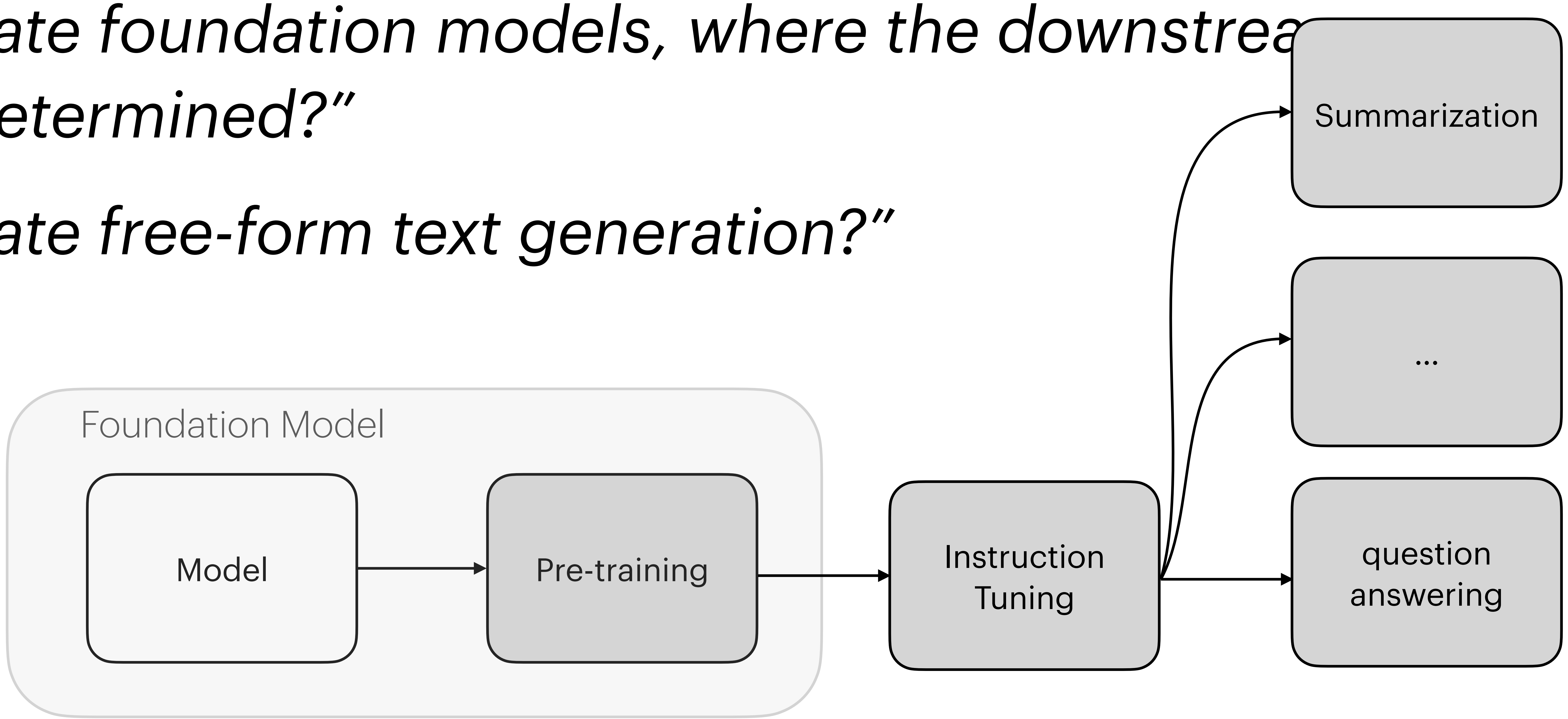
Recap: Text generation

- We have general purpose instruct-tuned models
  - Can perform zero-shot generalization
  - And can answer free-form questions



Recap: Text generation

- Central Questions
  - *"How do we evaluate foundation models, where the downstream case is yet to be determined?"*
  - *"How do we evaluate free-form text generation?"*



Evaluating General purpose systems: Benchmarks

- Evaluate model on a variety of tasks
  - Benchmarks — GLUE as an example
- GLUE claims to measure natural language understanding (NLU)
  - Consist 9 *diverse* tasks intended to measure language understanding

GLUE Tasks: Corpus of Linguistic Acceptability (CoLA)

- Single sentence task
- CoLA
- Metric: Mathhews correlation coefficient

clc95	0	*	In which way is Sandy very anxious to see if the students will be able to solve the homework problem?
c-05	1		The book was written by John.
c-05	0	*	Books were sent to each other by the students.
swb04	1		She voted for herself.
swb04	1		I saw that gas can explode.

GLUE Tasks: Microsoft Research Paraphrase Corpus (MRPC)

- Paraphrase/Similarity task
- Metric: F1 / Accuracy

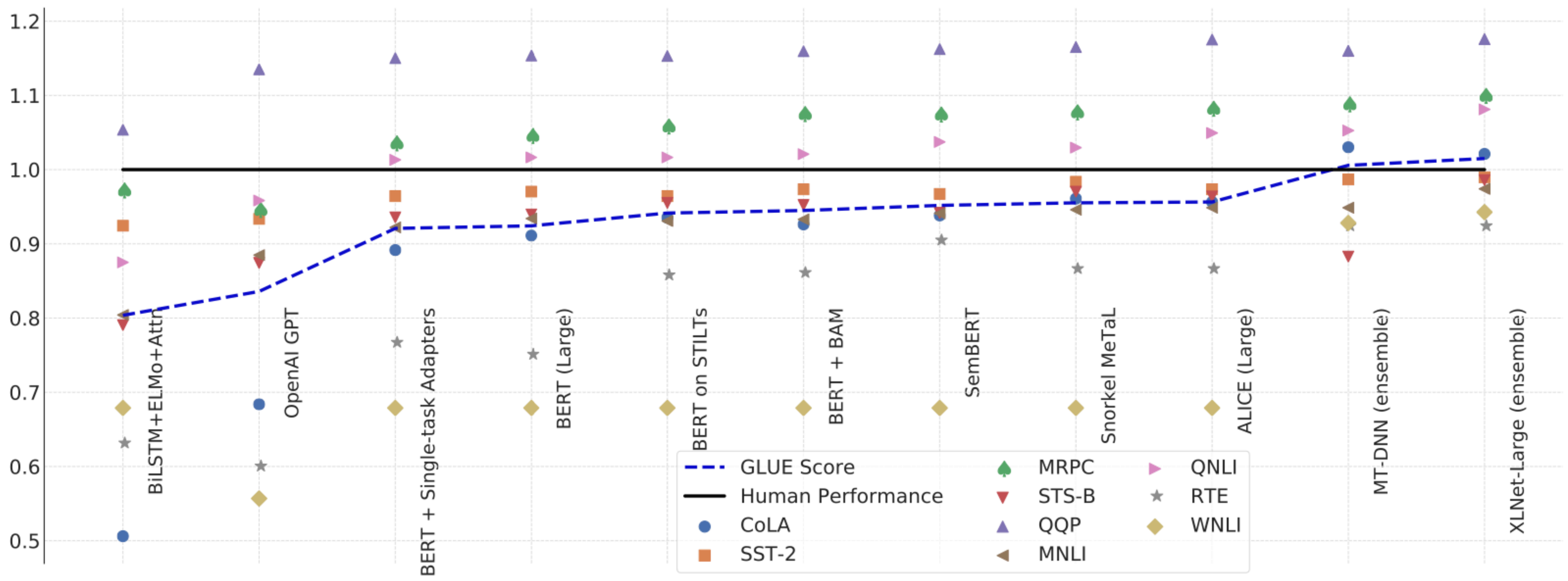
- **Sentence 1**: Amrozi accused his brother, whom he called "the witness", of deliberately distorting his evidence.
- **Sentence 2**: Referring to him as only "the witness", Amrozi accused his brother of deliberately distorting his evidence.
- **Class**: 1 (true paraphrase)

Benchmark Performance

Model	Avg	Single Sentence CoLA	SST-2	Similarity and Paraphrase MRPC	QQP	STS-B	Natural Language Inference MNLI	QNLI	RTE	WNLI
Single-task	64.8	**35.0**	90.2	68.8/80.2	**86.5/66.1**	55.5/52.5	**76.9/76.7**	61.1	50.4	**65.1**
Multi-task	**69.0**	18.9	**91.6**	**77.3/83.5**	85.3/63.3	72.8/71.1	75.6/75.9	81.7	**61.2**	**65.1**
CBoW	58.9	0.0	80.0	73.4/81.5	79.1/51.4	61.2/58.7	56.0/56.4	75.1	54.1	62.3
Skip-Thought	61.5	0.0	81.8	71.7/80.8	82.2/56.4	71.8/69.7	62.9/62.8	74.7	53.1	**65.1**
InferSent	64.7	4.5	85.1	74.1/81.2	81.7/59.1	75.9/75.3	66.1/65.7	79.8	58.0	**65.1**
DisSent	62.1	4.9	83.7	74.1/81.7	82.6/59.5	66.1/64.8	58.7/59.1	75.2	56.4	**65.1**
GenSen	66.6	7.7	83.1	76.6/83.0	82.9/59.8	**79.3/79.2**	71.4/71.3	**82.3**	59.2	**65.1**

See leaderboard here: https://gluebenchmark.com/leaderboard

Benchmark saturation



Problems with Aggregation

- Mean aggregation can be problematic

	Task 1	Task 2	Task 3	Task 4	Task 5	Mean
Model A	**0.2**	**0.15**	**0.2**	0.4	**0.3**	0.25
Model B	0.1	0.01	0.02	**0.9**	0.25	**0.256**

- Currently people are proposing different methods
  - e.g. borda rank
  - Or not providing an aggregate

Modern Benchmarks

- Example modern benchmark using the same approach
  - https://scandeval.com/

Qualitative Evaluations — the vibe check

- **Q**: How would we go about evaluating human responses?

Quantifying Qualitative Evaluations

- Thing we see in practice:
  - Numeric Scores:
    - Multiple choice (games, exams, ...)
    - Free form exam → quantify in a grade
  - Free-form Text
    - Reviews (e.g. on a book) — often use stars*
    - Written report
  - Oral feedback — Probably the highest quality for the recipient

* but stars is a fairly recent thing

Quantifying Qualitative Ratings

- Example leaderboard: https://lmarena.ai/

Who is margrete vestager

Margrethe Vestager is a Danish politician known for her significant role in European competition policy. Born on April 13, 1968, in Glostrup, Denmark, she earned a Master of Science in Economics from the University of Copenhagen in 1993. (Wikipedia)

Vestager's political career began in Denmark, where she served as Minister for Ecclesiastical Affairs and Minister of Education from 1998 to 2001. She later became the leader of the Danish Social Liberal Party and held the position of Minister for Economic and Interior Affairs from 2011 to 2014. (Wikipedia)

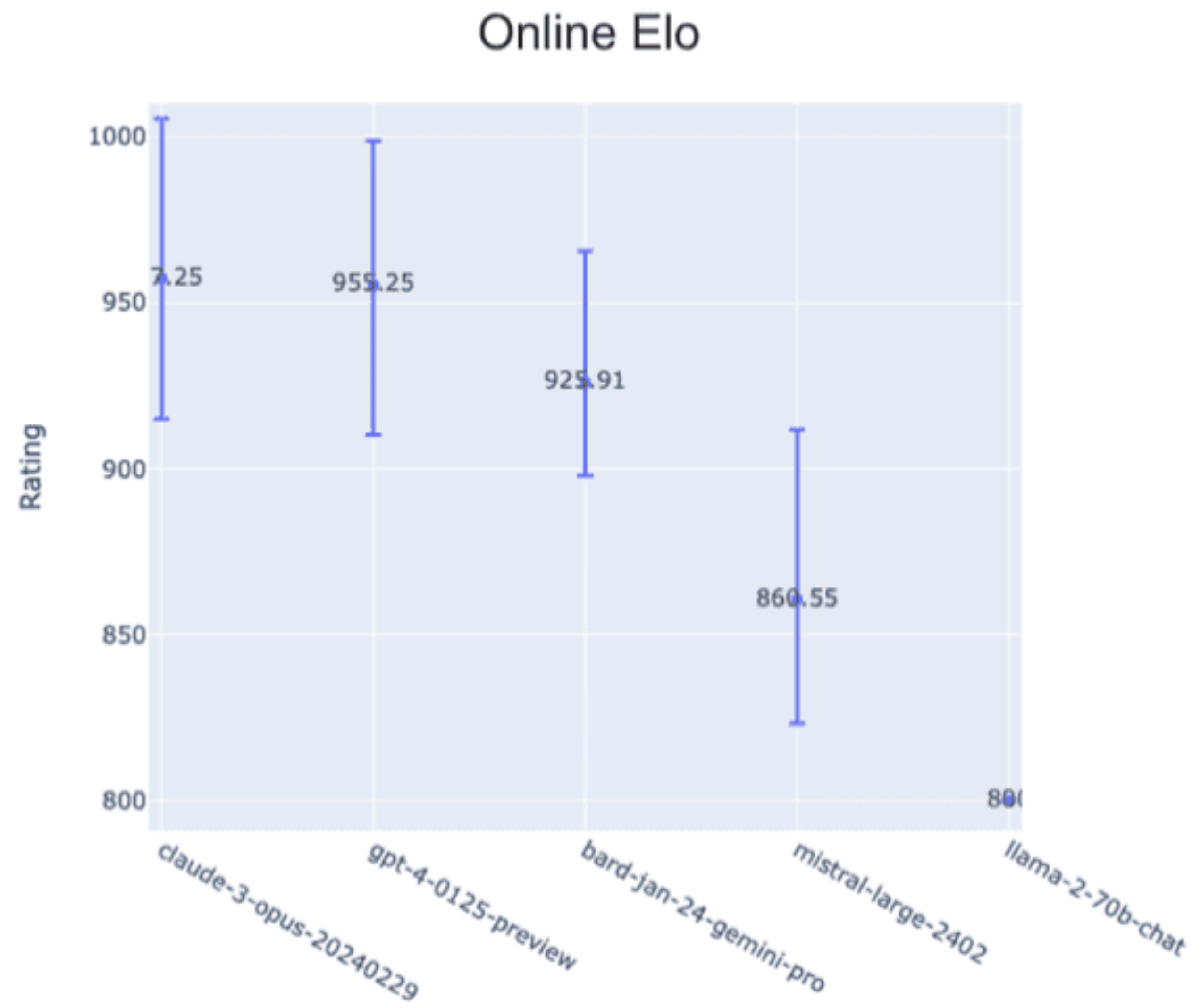
Who is Margrethe Vestager?

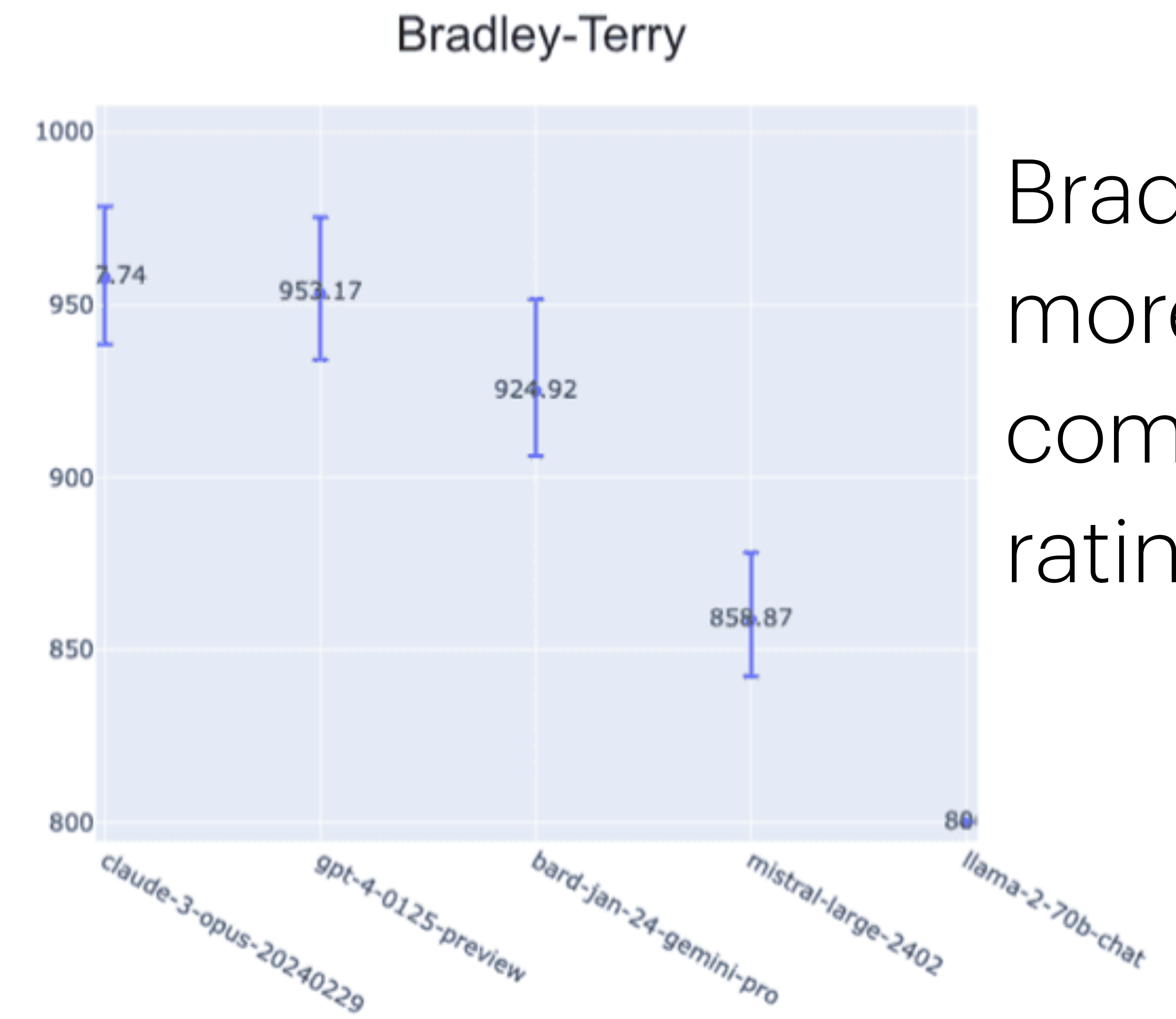
Margrethe Vestager is a Danish politician and European Commissioner known for her work in competition policy and technology regulation within the European Union. She has served as the Executive Vice President of the European Commission for a Europe Fit for the Digital Age and as Commissioner for Competition. Vestager gained prominence for her tough stance on antitrust issues, particularly in cases involving major technology companies such as Google, Apple, and Amazon, often leading to significant fines and regulatory changes.

Q: Which one do you prefer?

Quantifying pairwise comparisons

- Elo rating
  - e.g. known from chess
  - Computed without access to the history of games
  - It considers the posibility that the performance of a player changes over time
  - Overvalues recent wins
- Bradley Terry model
  - $P(m > m')$: "The probability that model *m* beats model *m'*"





Bradley-Terry provides more stable rating when compared to the Elo rating

Sources
& Notes

Further discussion and image source: https://hippocampus-garden.com/elo_vs_bt/

CENTER FOR HUMANITIES COMPUTING

LLM as a Judge

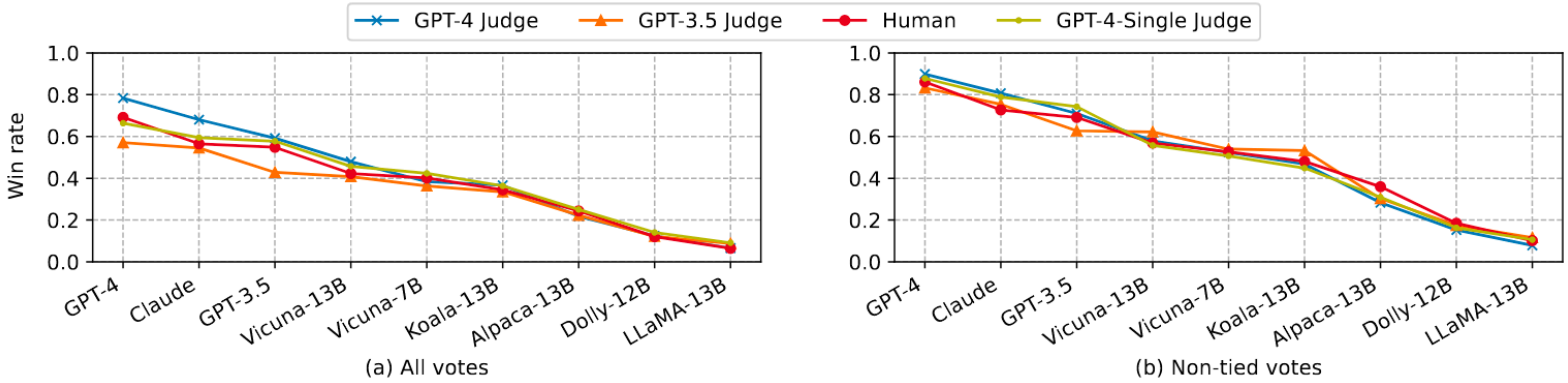
- Key limitations of qualitative evaluations
  - “Scalability”
- What about that reward model from last time?
- Can we use LLMs to evaluate our systems?

Sources & Notes

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena (arXiv:2306.05685). arXiv. https://doi.org/10.48550/arXiv.2306.05685

LLM as a Judge

- Achieve >80% agreement ~ humans



Sources & Notes

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena (arXiv:2306.05685). arXiv. https://doi.org/10.48550/arXiv.2306.05685

CENTER FOR HUMANITIES COMPUTING

Limitations of LLM Judges: Position bias

- Method:
  - Generating two responses from gpt 3.5
- Question does order influence preference
- Also seen in humans

Table 2: Position bias of different LLM judges. Consistency is the percentage of cases where a judge gives consistent results when swapping the order of two assistants. "Biased toward first" is the percentage of cases when a judge favors the first answer. "Error" indicates wrong output formats. The two largest numbers in each column are in bold.

Judge	Prompt	Consistency	Biased toward first	Biased toward second	Error
Claude-v1	default	23.8%	**75.0%**	0.0%	1.2%
GPT-3.5	default	46.2%	**50.0%**	1.2%	2.5%
GPT-4	default	**65.0%**	30.0%	5.0%	0.0%

Sources & Notes

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena (arXiv:2306.05685). arXiv. https://doi.org/10.48550/arXiv.2306.05685

Limitations of LLM Judges: Self enchancement Bias

- Does LLMs favor longer and more verbose responses?
- Method: Uses a “repetitive list” attack
  - Sample answers that contain lists
  - Rephrase list using gpt-4 and append to answer

Table 3: Failure rate under “repetitive list” attack for different LLM judges on 23 answers.

Judge	Claude-v1	GPT-3.5	GPT-4
Failure rate	91.3%	91.3%	8.7%

Limitations of LLM Judges: Self enchancement Bias

- Does LLM judges favor responses generated by themselves?
- Results:
  - GPT-4 favor itself with 10% winrate
  - Claude favors itself with a 25% winrate
  - GPT-3.5 does not favor itself

Limitations of LLM Judges

- Limitations on evaluating math and reasoning
- Probably additional indiscovered biases and limitations

Quantifying Qualitative Ratings

- **Q**: Is quantification needed?
  - A few places where you can think of is:
    - Book recommendations
    - Feedback on an essay
    - …

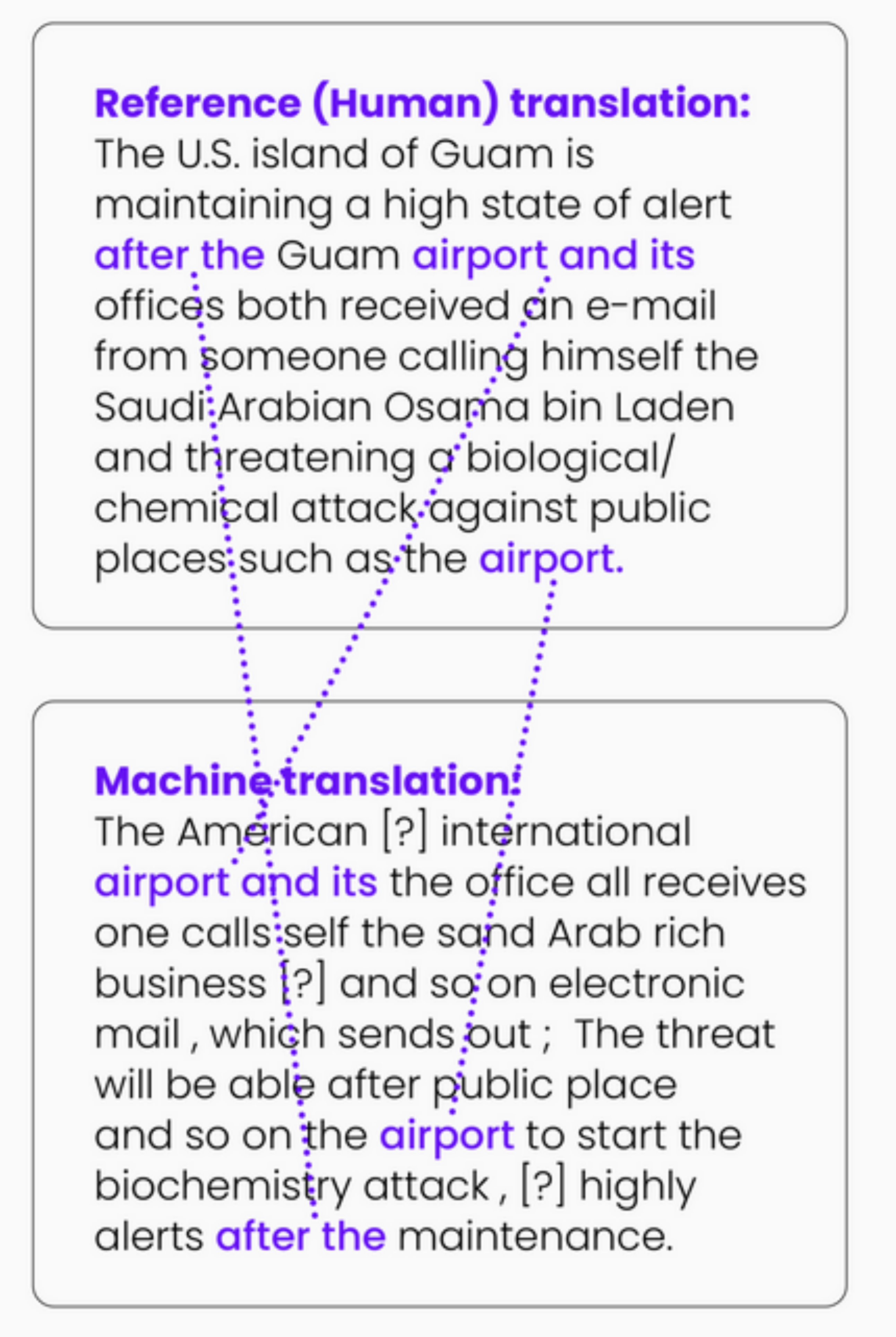
Recap: Qualitative Evaluations

- Many large open question with no clear answers
- We now have (NLP) tools to quantify qualitative evaluations
  - The boundary between these methods become less distinct
- Likely a many questions to tackle within this area

Quantitative Evaluations of text similarity

BLEU (2002): A method for automatic evaluation of machine Translation

- *“The closer a machine translation is to a professional human translation, the better it is”*
- What percentage of MT output **n-grams** can be found in the reference translation
- Between 0-1
  - 1: perfect lexical overlap
  - 0: no lexical overlap
- *Initially* high correlation with human preferences
- Similar to e.g. rouge (used for summarization)



Sources & Notes

https://aclanthology.org/P02-1040/

Image source: https://blog.modernmt.com/understanding-mt-quality-bleu-scores/

CENTER FOR HUMANITIES COMPUTING

BLEU in code

Many metrics are implemented in the “evaluate” library

```
>>> predictions = ["hello there general kenobi", "foo bar foobar"]
>>> references = [
...     ["hello there general kenobi", "hello there !"],
...     ["foo bar foobar"]
... ]
>>> bleu = evaluate.load("bleu")
>>> results = bleu.compute(predictions=predictions, references=references)
>>> print(results)
{'bleu': 1.0, 'precisions': [1.0, 1.0, 1.0, 1.0], 'brevity_penalty': 1.0, 'length_ratio': 1.1666666666666667, 'translation_length': 7, 'reference_length': 6}
```

BertScore

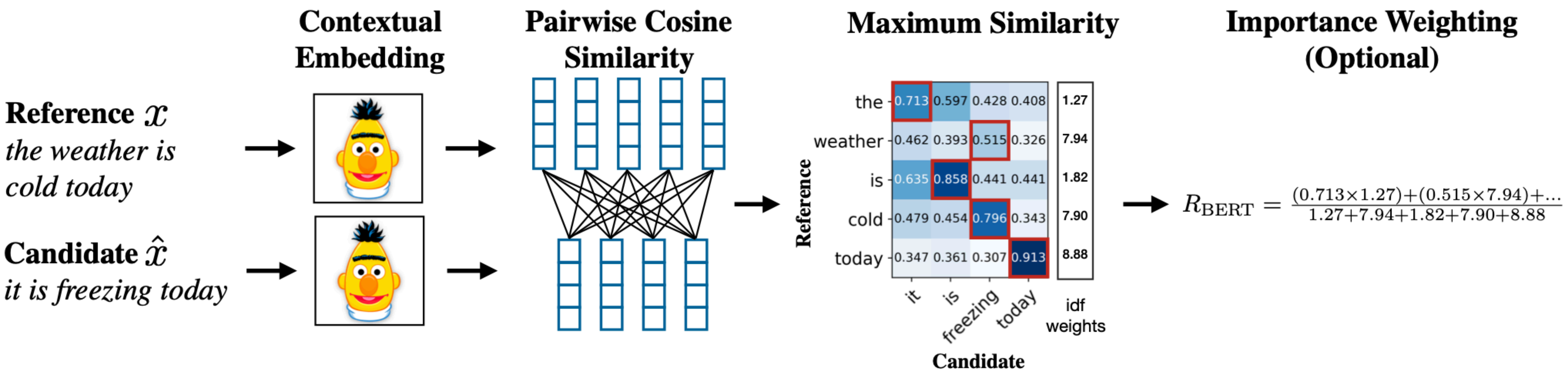


Figure 1: Illustration of the computation of the recall metric R_{BERT}. Given the reference x and candidate $\hat{x}$, we compute BERT embeddings and pairwise cosine similarity. We highlight the greedy matching in red, and include the optional idf importance weighting.

Image source: https://openreview.net/pdf?id=SkeHuCVFDr

BertScore in code

```
from evaluate import load
bertscore = load("bertscore")
predictions = ["hello world", "general kenobi"]
references = ["hello world", "general kenobi"]
results = bertscore.compute(predictions=predictions, references=references, model_type="distilbert-base-uncased")
print(results)
{'precision': [1.0, 1.0], 'recall': [1.0, 1.0], 'f1': [1.0, 1.0], 'hashcode': 'distilbert-base-uncased_L5_no-idf_version=0.3.10(hug_trans=4.10.3)'}
```

```
from evaluate import load
bertscore = load("bertscore")
predictions = ["hello world", "general kenobi"]
references = ["goodnight moon", "the sun is shining"]
results = bertscore.compute(predictions=predictions, references=references, model_type="distilbert-base-uncased")
print(results)
{'precision': [0.7380737066268921, 0.5584042072296143], 'recall': [0.7380737066268921, 0.5889028906822205], 'f1': [0.73
```

Sources & Notes

Source: https://huggingface.co/spaces/evaluate-metric/bertscore

Evaluation matter

- Example of where wrong evaluation lead to wrong conclusions
  - Emergent properties: “Suprising capabilities appear out of the blue”
  - Turns out it is a product of evaluation! (Or at least I don’t know of any strong evidence for emergence)
- Bullet 2
- Bullet 3

Error Analysis

- We have a system for answering user questions about our site, however we see that it is not always correct. We would like to know why and when it makes mistakes.
  - **Q**: How would you go about solving such a problem?

Error Analysis

- We have a system for answering user questions about our site, however we see that it is not always correct. We would like to know why and when it makes mistakes.
  - **Q**: How would you go about solving such a problem?
    - Find and examine cases of error
      - What are comonalities?
        - E.g. 80/100 errors are due to spelling errors
        - Or 61/100 errors are due are considered out of scope of the systems

Behavioural Testing

- Once we have figured out error group it might be ideal to test for it.
- We can do this using **behavioural testing**
- This allow us to test:
  - Minimum functionality test (MFT)
    - Simple test cases
  - invariance test (INV)
    - Certain pertubations should not change prediction
  - Direction expectation test (DIR)
    - If I do X I expect the probability to decrease/ increase

Test case	Expected	Predicted	Pass?
A Testing **Negation** with ***MFT*** — Labels: negative, positive, neutral			
Template: I {NEGATION} {POS_VERB} the {THING}.			
I can't say I recommend the food.	neg	pos	X
I didn't love the flight.	neg	neutral	X
...			
Failure rate = 76.4%			
B Testing **NER** with ***INV*** — Same pred. (inv) after removals / additions			
@AmericanAir thank you we got on a different flight to [Chicago → Dallas].	inv	pos neutral	X
@VirginAmerica I can't lose my luggage, moving to [Brazil → Turkey] soon, ugh.	inv	neutral neg	X
...			
Failure rate = 20.8%			
C Testing **Vocabulary** with ***DIR*** — Sentiment monotonic decreasing (↓)			
@AmericanAir service wasn't great. You are lame.	↓	neg neutral	X
@JetBlue why won't YOU help them?! Ugh. I dread you.	↓	neg neutral	X
...			
Failure rate = 34.6%			

Source: https://aclanthology.org/2020.acl-main.442.pdf