School of CIT –
Social Computing
Research Group

Technical
University
of Munich

# Advanced Natural Language Processing
# CIT4230002

Prof. Dr. Georg Groh
Miriam Anschütz, M.Sc.

# Lecture 4
# (Seq2Seq) Evaluation metrics

# Outline

- **Human evaluation**
- Types of seq2seq evaluation metrics
- Beyond evaluation
- Benchmarking LLMs

# Complexity of evaluation in NLP

- Evaluation in
  - Computer-vision

→ Calculate overlap

  - Classification

| Prediction: *Label 1* | Ground truth: *Label 0* |
|---|---|

  - Language generation

| Candidate: *I had a great weekend!* | Reference: *My weekend was superb!* |
|---|---|

?

- → Language is ambiguous and has a lot of nuances!

# Features of a good Seq2Seq output

- **Fluency**: Quality of target language text
  - Grammatically correct
  - Coherent
  - Considers language characteristics (e.g. syntax, lexical divergences)

- **Adequacy**: Preservation of exact meaning
  - No information missing
  - No repetition
  - No novel content

=> Hard to judge everything at once!

# Human evaluation

- **Human annotators**
  - Often online crowdworkers
  - Multilingual: rate criterion on defined scale
  - Monolingual: Compare to reference translation or rank pairs of candidate translations
  - Disagreement between raters → normalize and exclude outliers

  => Human evaluation is gold standard

- But:
  - Time-consuming
  - Expensive

=> Automated metrics for large-scale evaluations

# Discussion | Human evaluation - Subjectivity

- Task: Rate the candidate on a scale from 1 (best) to 5 (worst)

> Source: "*John became an older brother because Mary gave birth to a girl.*"
>
> Reference: *"Mary gave birth to a girl."*
>
> Candidate: *"John's mother had a baby."*

- Problems:
  - Candidate is very far from the reference… but correct!
    → show multiple references and source
  - Rating is subjective
    → Give objective criteria
    → Evaluate grammaticality, adequacy, .. independently
  - But: increases workload even more!

# Discussion | Human evaluation - Reproducibility

- **Reproducible** = Same results when running experiment/questionnaire multiple times

- Are human evaluations reproducible?

- Belz at al. (2023):
  *"Overall, we estimate that just 5% of human evaluations are repeatable in the sense that (i) there are no prohibitive barriers to repetition, and (ii) sufficient information about experimental design is publicly available for rerunning them. Our estimate goes up to about 20% when author help is sought."*

# Discussion | Human evaluation - Reproducibility

- Gold-standard evaluation, but no one can control/reproduce results?!
  - → Can authors just claim whatever they want?

- No, authors should report:
  - Study design
  - Demography of participants
  - Questionnaire
  - …

- **Don’t trust human evaluation scores unconditionally!**

# Outline

- Human evaluation
- **Types of seq2seq evaluation metrics**
- Beyond evaluation
- Benchmarking LLMs

# Features of a good evaluation metric

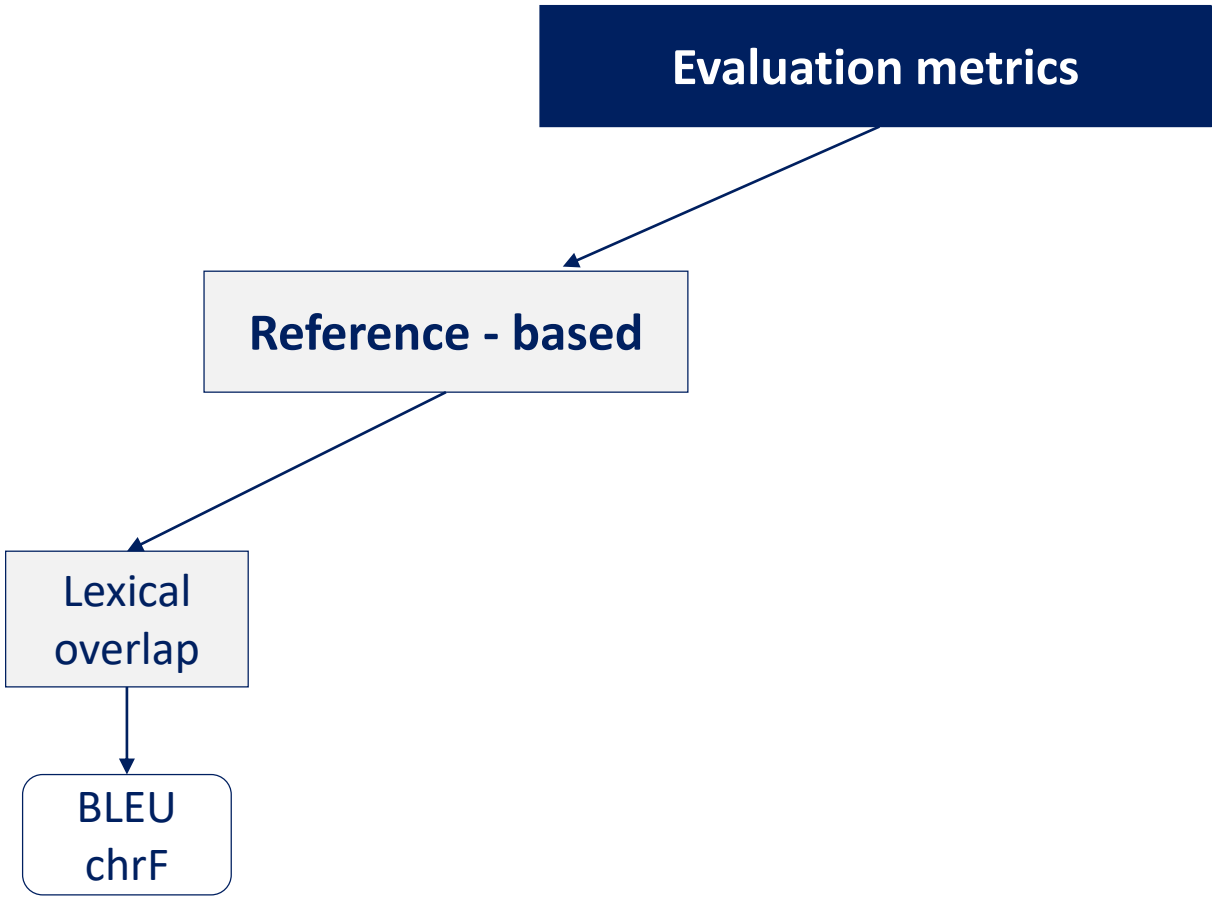
- Detects and punishes
  - Violations against adequacy
  - Violations against fluency
  - Copy-pasting input

- High correlation with human evaluations

| Evaluation | Sample 1 | Sample 2 | … | Correlation with human |
|---|---|---|---|---|
| Human | 25,1 | 37,4 | … | 100% |
| Metric 1 | 27,3 | 40,9 | … | High |
| Metric 2 | 15,6 | 12,3 | … | low |

- Fast and inexpensive to compute

# Types of evaluation metrics



# Lexical overlap with reference: BLEU

- "BLEU": BiLingual Evaluation Understudy
  - Rank candidates by word *n*-gram overlap
  - Purely precision based
- Calculates ratio of overlapping *n*-grams to total num. *n*-grams in candidate
- Weighted sum over different values for *n*
- Adds a brevity penalty for too short sentences
- Sensitive to tokenization (e.g. punctuation, compounds) and max *n*
  → standardized version is SacreBLEU

# Problems of overlap-based methods

- No synonym awareness

Reference: "*My weekend was superb*"
Candidate 1: "*My weekend was great*" → SacreBLEU 59.46
Candidate 2: "*My weekend was bad*" → SacreBLEU 59.46

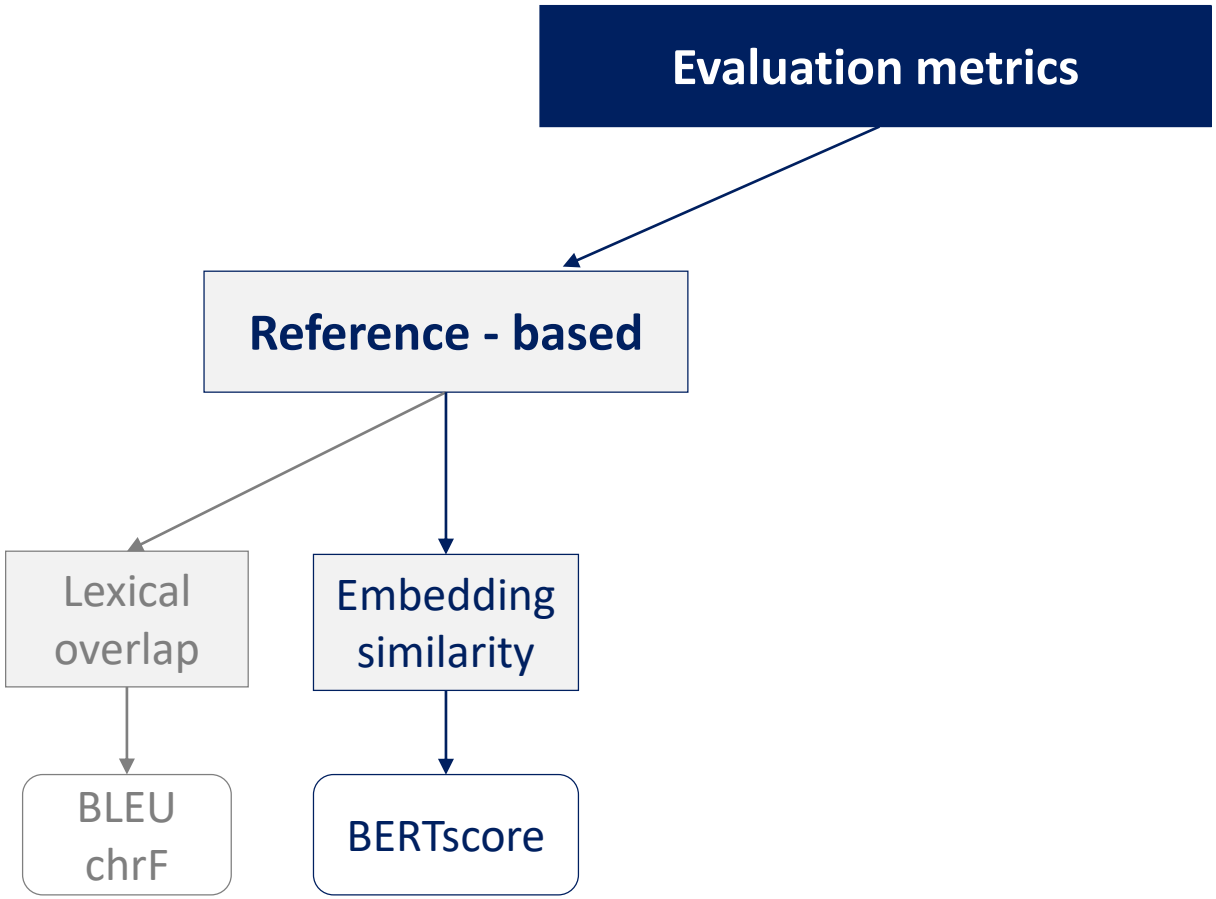
- Partly robust to changes in word/n-gram order

Reference: "*At the weekend, we visited my grandma's house and ate cake.*"
Candidate 1: "*At the weekend, we ate cake.*" → SacreBLEU 37.01
Candidate 2: "*At the weekend, we ate my grandma's house.*" → SacreBLEU 41.15

- No coverage of cross-sentence properties (SentenceBLEU does)

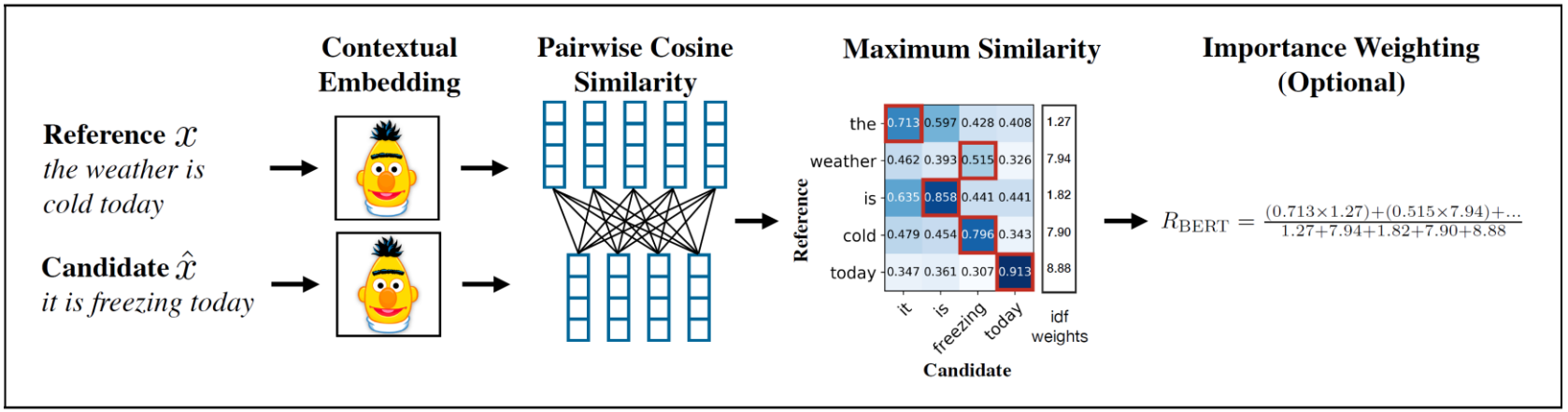
Reference: "*At the weekend, we visited my grandma's house and ate cake.*"
Candidate 1: "*At the weekend, we visited my grandma's house.*" → SacreBLEU 66.94
Candidate 2: "*At the weekend, we visited my grandma's house. And we ate cake.*" → SacreBLEU 64.75

# Types of evaluation metrics



# Recap | BERTScore

- Token-wise cosine similarity of embedding representations




- Tokens in $x$ matched to $\hat{x}$ → recall

$$R_{\text{BERT}} = \frac{1}{|x|}\sum_{x_i \in x} \max_{\tilde{x}_j \in \tilde{x}} x_i \cdot \tilde{x}_j$$

- Tokens in $\hat{x}$ matched to $x$ → precision

$$P_{\text{BERT}} = \frac{1}{|\tilde{x}|}\sum_{\tilde{x}_j \in \tilde{x}} \max_{x_i \in x} x_i \cdot \tilde{x}_j$$

# Problems/characteristics of embedding-based methods

- Latent representation not always distinctive (e.g. negation/antonyms)

| Reference: "*This house is in a big city.*" | SacreBLEU | BERTscore F1 |
|---|---|---|
| Candidate 1: *"The house is in a big city."* | 84,08 | 1.0 |
| Candidate 2: *"The house is not in a big city."* | 51.33 | 0.97 |

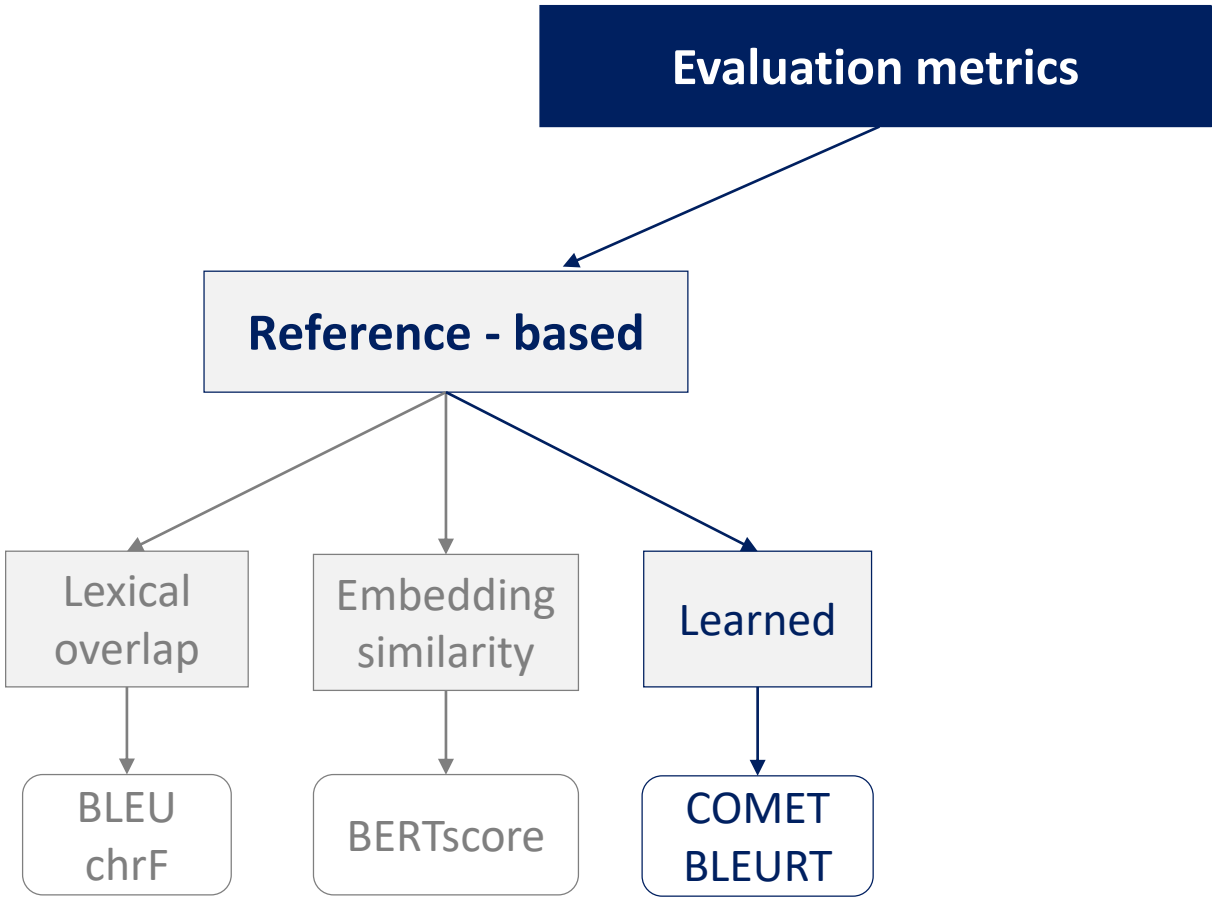
- Grammatical correctness not always covered (e.g. Yoda slang)

| Reference: "*This house is in a big city.*" | SacreBLEU | BERTscore F1 |
|---|---|---|
| Candidate 2: *"The house in a big city is."* | 39.76 | 0.95 |

- Performance dependent on model availability/language dependent

| Reference: "*Dieses Haus ist in einer großen Stadt.*" | SacreBLEU | BERTscore F1 |
|---|---|---|
| Candidate 2: *"Das Haus in einer großen Stadt ist."* | 39.76 | 0.93 |

# Types of evaluation metrics



# Annually Shared Task on MT metrics

## WMT23 Metrics Task

This shared task will examine automatic evaluation metrics for machine translation.

### Task Description

We will provide you with the source sentences, output of machine translation systems and reference translations.

1. Official results: Correlation with MQM scores at the sentence and system level for the following language pairs:
   - Hebrew-English **(NEW!)**
   - Chinese-English
   - English-German **This will be a paragraph-level task!**
2. Secondary Evaluation: Correlation with official WMT Human Evaluation at the sentence and system level.

### Subtasks:

1. QE as a Metric: In this subtask participants have to score machine translation systems without access to reference translations
2. Challenge Sets: While other participants are worried with building stronger and better metrics, participants of this subtask have to build challengesets that identify where metrics fail!

| **Reference *x*** | **Candidate *x̂*** | **Human score *y*** |
|---|---|---|
| My weekend was superb | My weekend was great | 1.0 |

# Learned metrics: BLEURT

- "BLEURT": Bilingual Evaluation Understudy with Representations from Transformers
  - Pre-trained BERT with linear layer for score prediction
  - Trained on human scores
  - Additional pre-training on synthetic data

- Synthetic training data generated with:
  - Mask-filling with BERT model
  - Backtranslation
  - Randomly dropping words

# Learned metrics: COMET

- "COMET": Cross-lingual Optimized Metrics for Evaluation of Translations
  - Pre-trained multilingual embedding (XLM-RoBERTa) with linear layer for score prediction
  - Trained on human scores

- Many alternative versions:
  - Translation ranking objective
  - Ensemble model with sequence tagging model
  - COMETKiwi: reference-free
  - XCOMET: explainable score with error span annotation

# Learned metrics: Alternative approaches

- MaTESe: Evaluation as error span detection
  - Sequence tagging to predict mistakes
  - Predicting error severity
- Prism: Evaluation as paraphrase generation task
  - Train model for paraphrase generation
  - Estimate paraphrase distance between reference/source and candidate
- More recent/larger language models than BERT
- Ensembles of models
- …

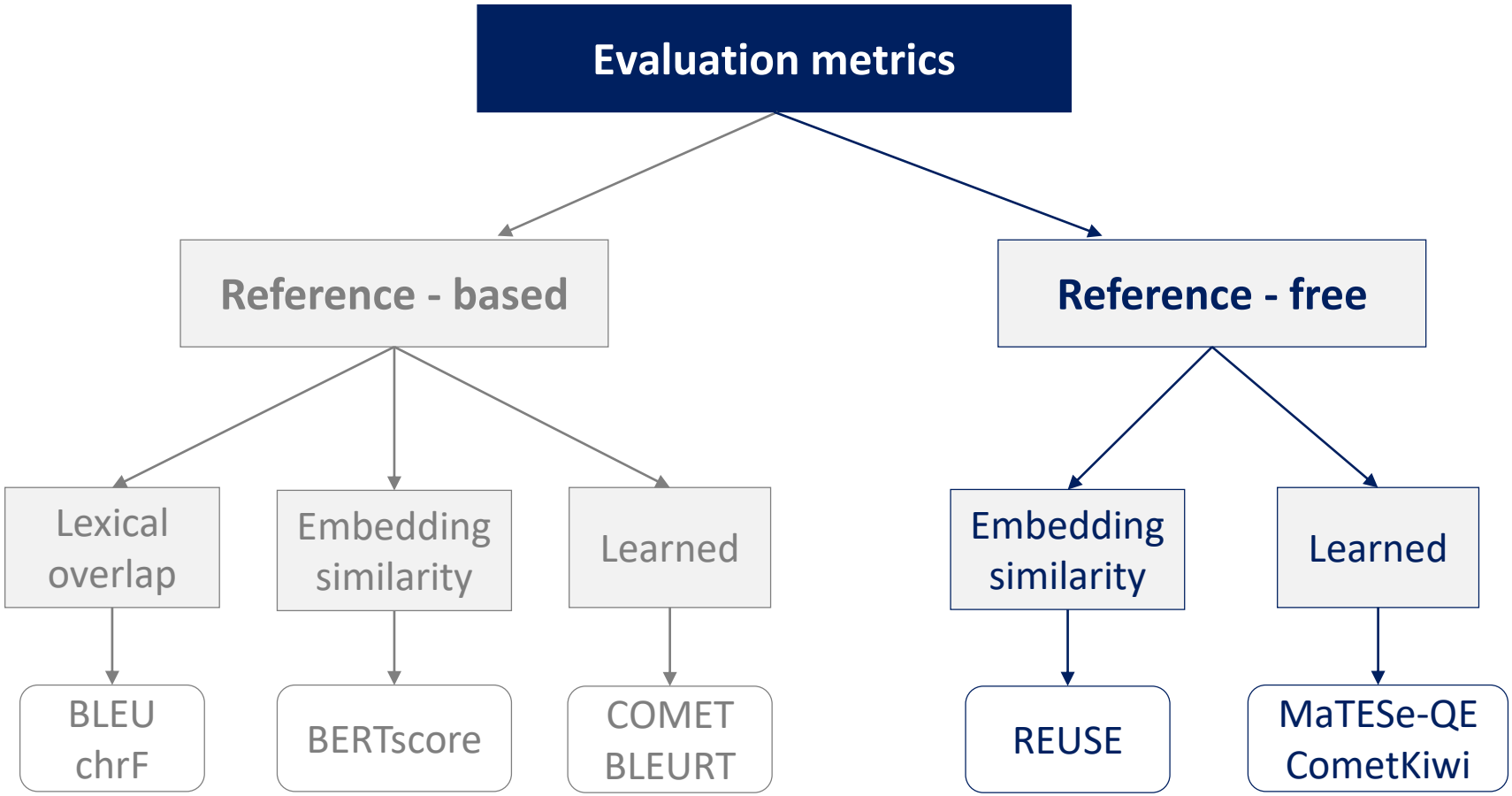
# Learned metrics: Discussion

- Best correlation with human scores (for the moment)
- Optimizable for specific tasks or criteria

But:

- Maaany different solutions
  - No one-fits-all solution
  - Harder to compare among papers
- Possible domain mismatch between training and evaluation data
- Still use a reference..

# Types of evaluation metrics



# From reference to source comparison

- Idea: Compare semantic overlap against source instead of reference
  - Example: COMET to CometKiwi
- Embedding similarity: Exploit multilingual embeddings
- Learned:
  - Multilingual base models
  - Similar learning objectives as reference-based

→ 2023 WMT Metrics shared task: 4 of the top-7 metrics are reference-free!

→ Because of poor reference quality?

→ No more references in (close) future?

# Reference-free: Evaluating with LLMs

- Idea: Exploit LLMs for evaluation
  - No reference needed
  - Zero-/Few-shot prompts

- "GEMBA-MQM": Detecting Translation Quality Error Spans with GPT-4

(4)

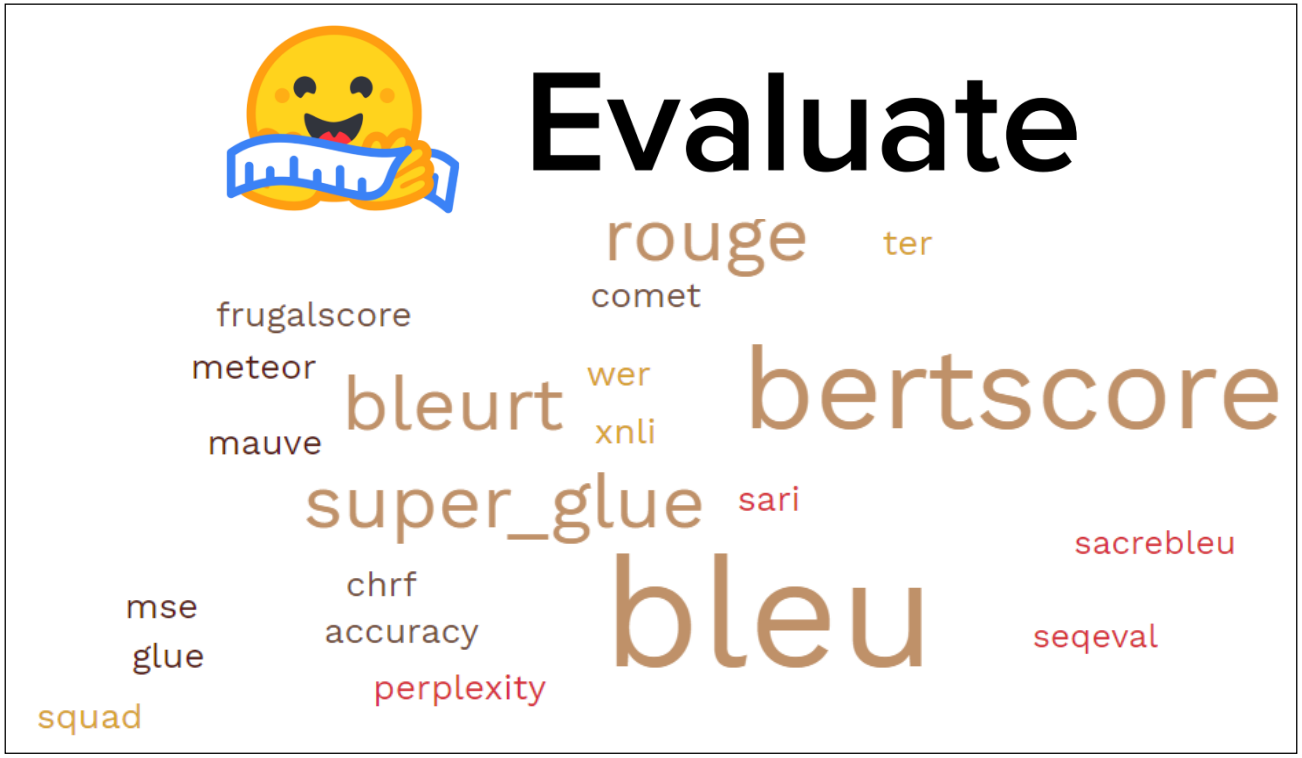
```
(System) You are an annotator for the quality of machine translation. Your task is to identify
errors and assess the quality of the translation.

(user) {source_language} source:\n
```{source_segment}```\n
{target_language} translation:\n
```{target_segment}```\n
\n
Based on the source segment and machine translation surrounded with triple backticks, identify
error types in the translation and classify them. The categories of errors are: accuracy
(addition, mistranslation, omission, untranslated text), fluency (character encoding, grammar,
inconsistency, punctuation, register, spelling),
locale convention (currency, date, name, telephone, or time format)
style (awkward), terminology (inappropriate  for context, inconsistent use), non-translation,
other, or no-error.\n
Each error is classified as one of three categories: critical, major, and minor.
Critical errors inhibit comprehension of the text. Major errors disrupt the flow, but what
the text is trying to say is still understandable. Minor errors are technically errors,
but do not disrupt the flow or hinder comprehension.

(assistant) {observed error classes}
```

Figure 1: The general prompt for GEMBA-MQM omits the gray part which performed subpar on internal data (we include it in GEMBA-locale-MQM). The "(user)" and "(assistant)" section is repeated for each few-shot example.

# Intermediate Summary

- Different ideas and benefits
- No one-fits-all solution but many work well in standard cases!

- Many metrics implemented in Huggingface Evaluate



# Outline

- Human evaluation
- Types of seq2seq evaluation metrics
- **Beyond evaluation**
- Benchmarking LLMs

# Faithfulness detection in “traditional” metrics

- These metrics don’t check for hallucinations
- Hallucinations change word order / structure of sentence

→ Hallucination detection works okay

| Reference: “*This house is in a big city.*” | SacreBLEU | BERTscore F1 | BLEURT |
|---|---|---|---|
| Candidate 1: “*This house is in a big city.*” | 100 | 1.0 | 1.0 |
| Candidate 2: “*This house is in the big city close to the ocean.*” | 26.20 | 0.97 | 0.0 |

# Explaining the score

- Idea: Highlight the wrong words in candidate

**Source:** (5)

This year's trend for a second Christmas tree in the bedroom sends sales of smaller spruces soaring

**Translation:**

Der diesjährige Trend für einen zweiten Weihnachtsbaum in der Schlafzimmer sendet Umsatz von kleineren Fichten steigen

severity: Major — category: Grammar

severity: Major — category: Mistranslation

→ Interpretable assessment for humans

→ Quality control of *metrics* (Do they punish the right things?)

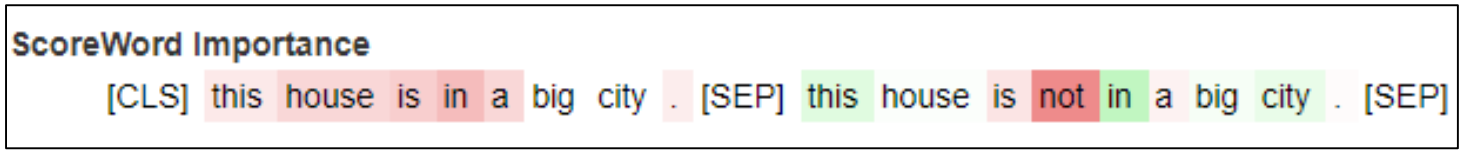
→ Quality control of *models* (Where do they fail?)

→ Post-hoc output correction

# How to explain?

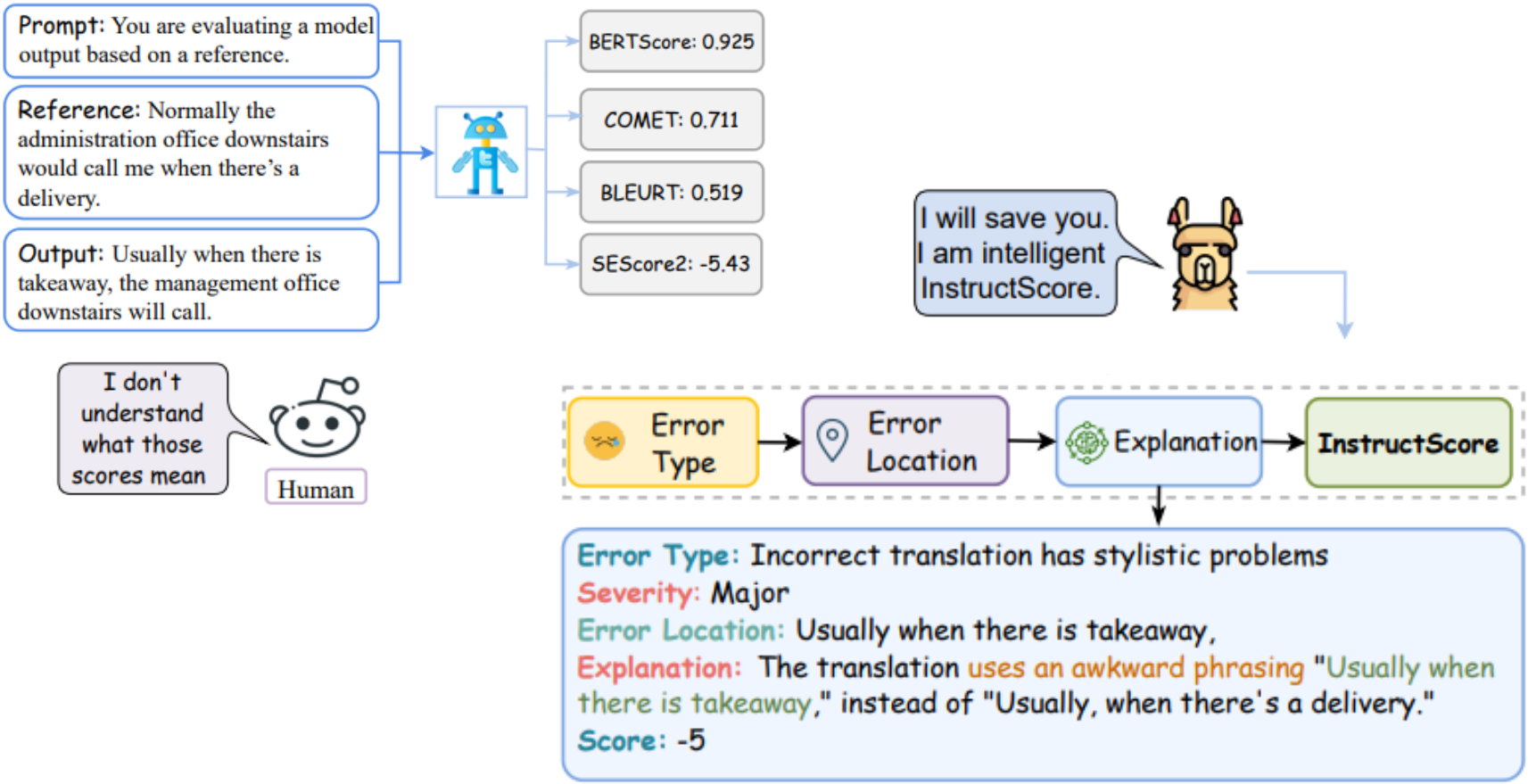
- Jointly learning sentence level score and word level annotations
- Post-hoc annotation of errors on word level
  - Sequence tagging models to identify erroneous spans
  - Grammatical error correction models to predict type of error
- Model-agnostic explainability frameworks to identify words influencing overall score (e.g. LIME)
- By-design interpretable metrics (e.g. edit-based scoring)

Explaining BLEURT model output:



# Explainable score with LLMs

- INSTRUCTSCORE: Explainable score based on LLaMA
  - Fine-tuned with GPT-3 and GPT-4
  - Evaluates different aspects simultaneously

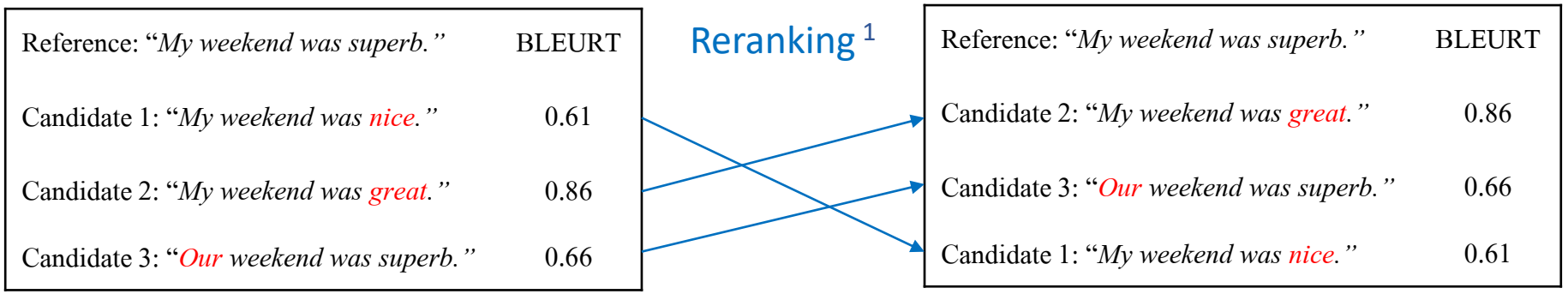



adapted from (8)

# Candidate ranking

- Problem: models often overly certain of output
  - Varying output by different decoding strategies

- Idea: Candidate ranking
  - Generate multiple translation candidates
  - Rank with evaluation metric
  - Return highest-ranking candidate




[1] In production setting no reference available → use reference-free metrics

# Outline

- Human evaluation
- Types of seq2seq evaluation metrics
- Beyond evaluation
- **Benchmarking LLMs**

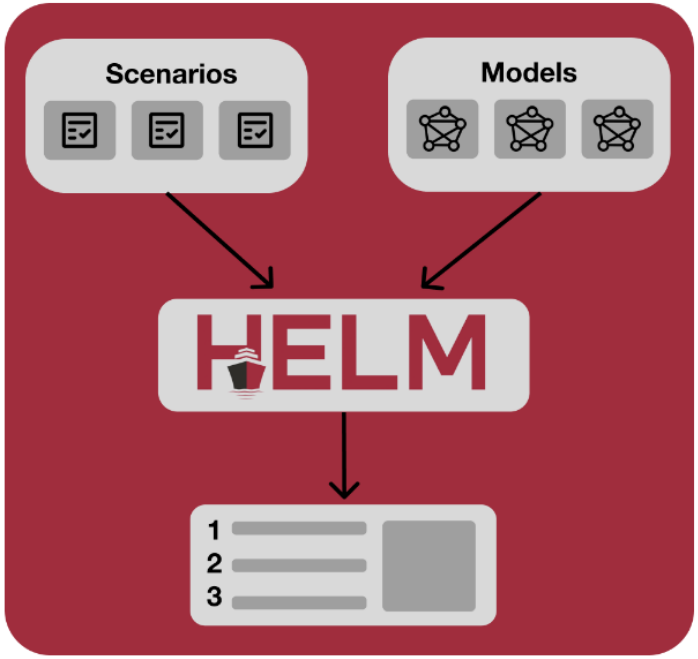
# Benchmarking

- LLMs perform more than one single task
  → Evaluation needs to be broader

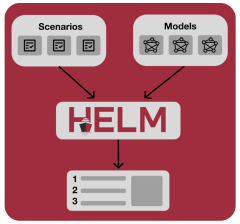
- Benchmarking on task collections

SuperGLUE

AlpacaEval Leaderboard



# Holistic Evaluation of Language Models (HELM)




- Benchmarking models on 87 scenarios, like ..
  - **Question answering** (MMLU, TruthfulQA, ..)
  - **Summarization** (CNN/DailyMail, XSUM)
  - **Sentiment analysis** (IMDB)
  - **Reasoning** (HumanEval for Code, MATH, GSM8K, ..)
  - ...

As of 3.6.24

## HELM Leaderboard

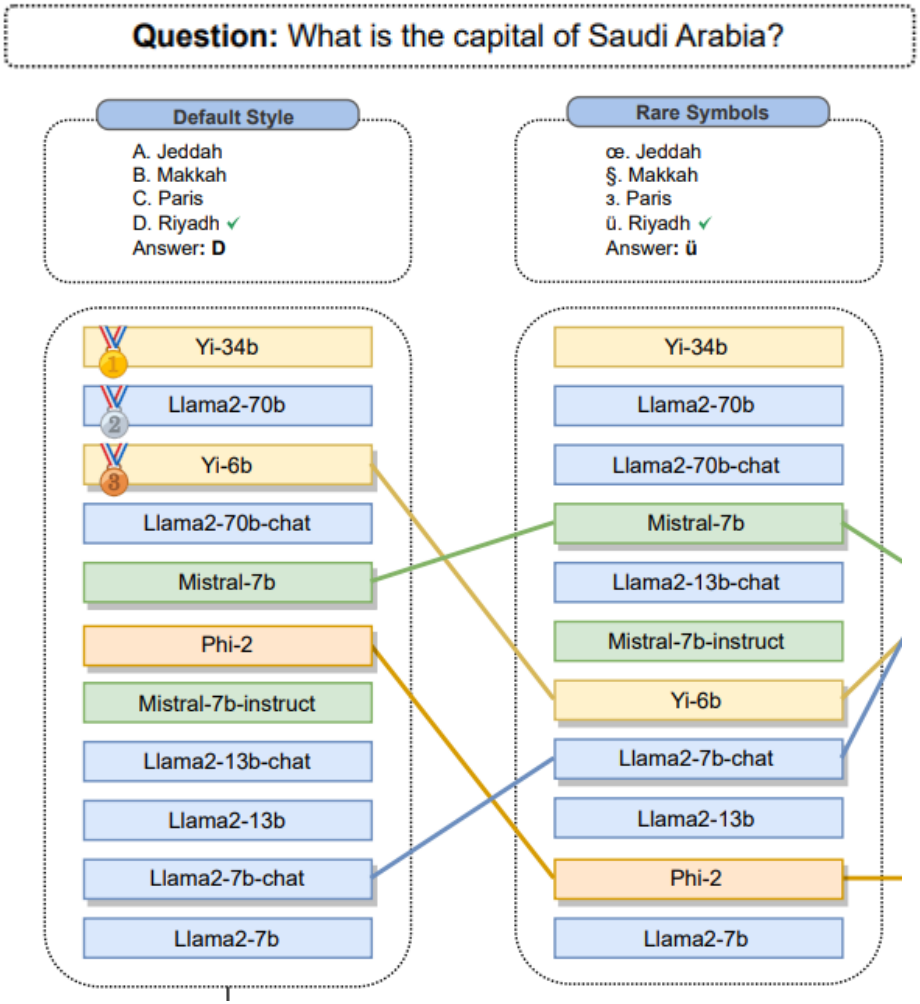
The HELM leaderboard shows how the various models perform across different scenarios and metrics.

Accuracy | Efficiency | General information

| Model | Mean win rate | NarrativeQA - F1 | NaturalQuestions (open) - F1 | NaturalQuestions (closed) - F1 |
|---|---|---|---|---|
| GPT-4 (0613) | **0.957** | 0.768 | 0.79 | 0.457 |
| Llama 3 (70B) | 0.902 | **0.798** | 0.743 | 0.475 |
| Mixtral (8x22B) | 0.855 | 0.779 | 0.726 | **0.478** |
| Palmyra X V3 (72B) | 0.826 | 0.706 | 0.685 | 0.407 |
| GPT-4 Turbo (1106 preview) | 0.821 | 0.727 | 0.763 | 0.435 |
| PaLM-2 (Unicorn) | 0.781 | 0.583 | 0.674 | 0.435 |

# Discussion: Prompt sensitivity

- LLMs are evaluated in a zero-/few-shot manner
- Benchmark performance sensitive to experimental setup[1]:




→ Changing the answer identifier changes the models' ranks!

1 Alzahrani et al. 2024. Prompt sensitivity in LLM leaderboards (https://arxiv.org/abs/2402.01781)

# Discussion: Data contamination

- Situation:
  - Benchmark data is open-source
  - LLMs are trained on internet data
  - → LLMs may have used benchmarks in training!

- Some models can reproduce samples from benchmarks like GSM8k[1]
- Models perform better on older benchmarks[2]

1 Xu et al. 2024. Benchmarking Benchmark Leakage in LLMs (https://arxiv.org/abs/2404.18824)
2 Roberts et al. 2023. Data Contamination over time (https://arxiv.org/abs/2310.10628)

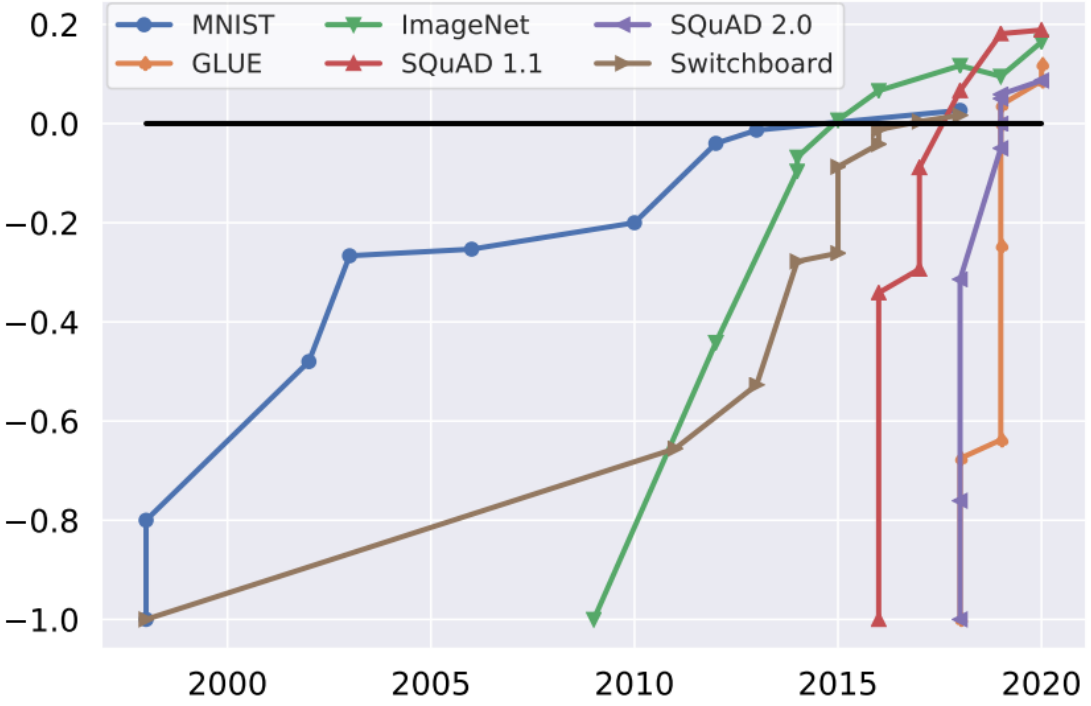
# Discussion: Data contamination

- Ideas to overcome data contamination:
  - Encrypting test data → automatic parsing no longer possible
  - Only evaluate on benchmarks younger than the models
  - Contamination detectors

- Benchmarking closed-source models leaks the benchmarks to them!
- → Is zero-shot benchmarking even possible?

# Discussion: Saturation

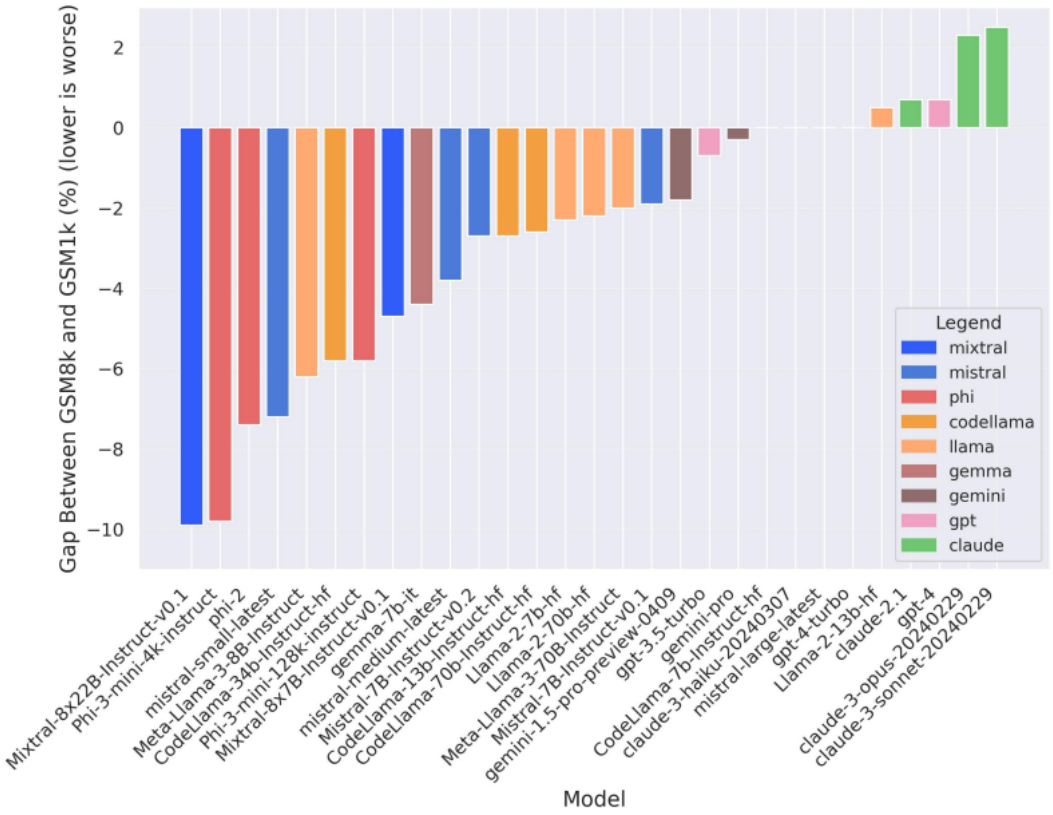
- Kiela et al. (2021)[1]:
  - Black line = human performance
  - → Models outperform humans
  - → Models "saturate" the benchmarks within a few years




1 Kiela et al. 2021. Dynabench: Rethinking Benchmarking in NLP (https://aclanthology.org/2021.naacl-main.324/)

# Discussion: Overfitting

- Models like Phi or Mistral show systematic overfitting on GSM8k[1]




Notable models arranged by their drop in performance between GSM8k and GSM1k (lower is worse). We notice that Mistral and Phi top the list of overfit models, with almost 10% drops on GSM1k compared to GSM8k, while models such as Gemini, GPT, and Claude show little to no signs of overfitting.

1 Zhang et al. 2024. Model performance on GSM8k grade school arithmentic (https://arxiv.org/abs/2405.00332)

# Discussion: Overfitting

- Situation:
  - Models train on synthetic data, generated by GPT4
  - Models evaluate with GPT4-judge

- → Models overfit to GPT4 instead of real-world problems

# Further discussion

- Benchmarks mostly available in English
- Multilinguality mostly for translation

- Multilingual benchmarks exist!
  - **MEGA**: Multilingual Evaluation of Generative AI
    - 16 datasets, 70 languages
  - **GlobalBench**:
    - 966 datasets in 190 languages.
  - **XTREME**: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization
    - 9 tasks, 40 languages

# Evaluating LLMs in human interaction

- Idea: Usability evaluation with human actors
- Chatbot Arena[1]:
  - Evaluate based on human preferences
  - Pairwise ranking → less task ambiguity and subjectivity

## LMSYS Chatbot Arena Leaderboard

| Model | Arena Elo | 95% CI | Votes | Organization | License | Knowledge Cutoff |
|---|---|---|---|---|---|---|
| GPT-4o-2024-05-13 | 1287 | +4/-4 | 32181 | OpenAI | Proprietary | 2023/10 |
| Gemini-1.5-Pro-API-0514 | 1267 | +5/-4 | 25519 | Google | Proprietary | 2023/11 |
| Gemini-Advanced-0514 | 1266 | +5/-5 | 27225 | Google | Proprietary | Online |
| Gemini-1.5-Pro-API-0409-Preview | 1257 | +3/-3 | 55731 | Google | Proprietary | 2023/11 |

1 Chiang et al. 2024. Chatbot Arena - Evaluating LLMs by Human Preference (https://arxiv.org/abs/2403.04132)

# Final thoughts

- Different ideas to evaluate candidates against reference or source
- Highlighting mistakes in output with metric explainability
- Be careful with human evaluation / benchmarks!
  - Results are fragile and easy to alter
  - All details on experimental setup should be reported
- Model evaluation criterion beyond output quality:
  - Induced ethical Biases
  - Reproducibility of results
  - Computational requirements and environmental impact
  - Vulnerability to adversarial use-cases
  - …

# Bibliography


(1) Chauhan et al. 2022. Machine Translation Evaluation Metrics Survey

(2) Sai et al. 2022. NLG Evaluation Metric Survey

(3) Belz et al. 2023. Reproducibility Crises in Human NLP Evaluation

(4) Kocmi and Federmann. 2023. Translation Error Detection with GPT-4

(5) Zerva et al. 2022. WMT shared task on quality estimation

(6) Freitag et al. 2023. WMT Metric Shared Task Findings

(7) Rei et al. 2022. COMET-22 descriptions

(8) Xu et al. 2023. INSTRUCTSCORE as explainable LLM score

(9) Sebastian Ruder. 2024. Blogpost on LLM evaluation

(10) Yann Dubois. 2024. Stanford CS224N Lecture on Benchmarking and Evaluation

# Study Approach

## Minimal

- work with the slides

## Standard

- minimal approach + read reference 9

## In-Depth

- standard approach + skim references 1, 2 and 10

See you next time!