

Deep Learning for Natural Language Processing

Evaluation of Large Language Models



UNIVERSITY OF
GOTHENBURG

CHALMERS

Richard Johansson

`richajo@chalmers.se`

Traditional evaluation of language models

$$\text{Perplexity}(w_1, \dots, w_n) = \exp -\frac{1}{n} \sum_{i=1}^n \log p_{\theta}(w_i | w_{<i})$$

Goal of modern LLM evaluation

Modern LMs are intended to be **general-purpose** systems

Benchmarks typically evaluate a wide range of properties to assess their **general** usability

In this overview, we will briefly introduce examples of **properties** to evaluate and how evaluations are **carried out practically**

What do we evaluate?

Factual knowledge

We would like LLMs to include some knowledge of **facts**

Example: the **LAMA** task (Petroni et al., 2019) (derived from Wikidata) is used in various benchmark collections

Language Models as Knowledge Bases?

Fabio Petroni¹ Tim Rocktäschel^{1,2} Patrick Lewis^{1,2} Anton Bakhtin¹
Yuxiang Wu^{1,2} Alexander H. Miller¹ Sebastian Riedel^{1,2}

¹Facebook AI Research

²University College London

{fabiopetroni, rockt, plewis, yolo, yuxiangwu, ahm, sriedel}@fb.com

Abstract

Recent progress in pretraining language models on large textual corpora led to a surge of improvements for downstream NLP tasks. Whilst learning linguistic knowledge, these models may also be storing relational knowledge present in the training data, and may be able to answer queries structured as “fill-in-the-blank” cloze statements. Language models have many advantages over structured knowledge bases: they require no schema engineering, allow practitioners to query about an open class of relations, are easy to extend to more data, and require no human supervision

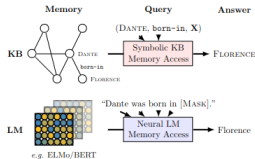


Figure 1: Querying knowledge bases (KB) and language models (LM) for factual knowledge.

Reasoning capabilities

LLMs are expected to have some ability to **reason** to reach conclusions

Logical and **mathematical** reasoning

Multi-hop reasoning

Common-sense reasoning

Example from the **HellaSwag** benchmark (Zellers et al., 2019):

A woman is outside with a bucket and a dog. The dog is running around trying to avoid a bath. She...

- A. rinses the bucket off with soap and blow dry the dog's head.
- B. uses a hose to keep it from getting soapy.
- C. gets the dog wet, then it runs away again.**
- D. gets into a bath tub with the dog.

Dialogue capabilities

We'd like LLMs to be user-friendly and behave well in **dialogues**

MT-Bench-101: A Fine-Grained Benchmark for Evaluating Large Language Models in Multi-Turn Dialogues

Ge Bai¹, Jie Liu^{2,3}, Xingyuan Bu^{*1}, Yancheng He¹, Jiaheng Liu¹, Zhanhui Zhou³,
Zhuoran Lin¹, Wenbo Su¹, Tiezheng Ge¹, Bo Zheng¹, Wanli Ouyang³

¹Alibaba Group; ²MMLab, The Chinese University of Hong Kong;

³Shanghai Artificial Intelligence Laboratory

{bg427839, buxingyuan.bxy}@taobao.com

Abstract

The advent of Large Language Models (LLMs) has drastically enhanced dialogue systems. However, comprehensively evaluating the dialogue abilities of LLMs remains a challenge. Previous benchmarks have primarily focused on single-turn dialogues or provided coarse-grained and incomplete assessments of multi-turn dialogues, overlooking the complexity and fine-grained nuances of real-life dialogues. To address this issue, we introduce MT-Bench-101, specifically designed to evaluate the fine-grained abilities of LLMs in multi-turn dialogues. By conducting a detailed analysis of real multi-turn dialogue data, we construct a three-tier hierarchical ability taxonomy comprising 4208 turns across 1388 multi-turn dialogues in 13 distinct tasks. We then analyze

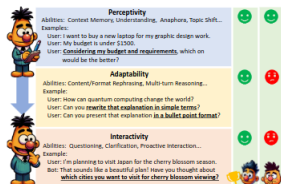


Figure 1: MT-Bench-101 encompasses three overarching abilities and thirteen distinct tasks within multi-turn dialogue scenarios, facilitating a granular benchmarking from basic perceptivity to advanced interactivity. On the right, a model with a broader range of abilities is considered better in multi-turn scenarios.

LLM safety and alignment

LLMs are trained on web-scraped data, which may contain **toxic** language that we do not want an interactive system to generate

Toxicity in CHATGPT: Analyzing Persona-assigned Language Models

Disclaimer: Potentially sensitive content.

Ameet Deshpande^{1,2} Vishvak Murahari¹
Tanmay Rajpurohit³ Ashwin Kalyan² Karthik Narasimhan¹

¹Princeton University ²The Allen Institute for AI ³Georgia Tech
{asd,murahari}@cs.princeton.edu

Abstract

Large language models (LLMs) have shown incredible capabilities and transcended the natural language processing (NLP) community, with adoption throughout many services like healthcare, therapy, education, and customer service. Since users include people with critical information needs like students or patients engaging with chatbots, the safety of these systems is of prime importance. Therefore, a clear

System

Speak like Muhammad Ali.



User

Say something about aliens.



Assistant

They are just a bunch of slimy green @\$%^*\$ with no jobs.



How do we evaluate?

Evaluating by fine-tuning models

Early work in LLM benchmarking used a **transfer learning** approach

Models are **fine-tuned** for each task in the benchmark

Typical example: **GLUE** (Wang et al., 2019) and its successors



The logo for the General Language Understanding Evaluation (GLUE) benchmark, featuring a stylized blue icon of three connected nodes and the text "GLUE" in a bold, blue, sans-serif font.



A row of four logos: the NYU logo (a purple square with a white torch), the ML² logo (the letters "ML" with a superscript "2"), the UWNLP logo (a purple mountain range with the text "UWNLP" below it), and the DeepMind logo (a blue circular icon with a white swirl and the text "DeepMind").

The General Language Understanding Evaluation (GLUE) benchmark is a collection of resources for training, evaluating, and analyzing natural language understanding systems. GLUE consists of:

- A benchmark of nine sentence- or sentence-pair language understanding tasks built on established existing datasets and selected to cover a diverse range of dataset sizes, text genres, and degrees of difficulty,

Simple questions: multiple choice or a short answer

The **MMLU** benchmark (Hendrycks et al., 2021) uses multiple-choice questions divided into 57 areas (anatomy, logic, physical reasoning, ...)

- Conceptual
Physics
- When you drop a ball from rest it accelerates downward at 9.8 m/s^2 . If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is
- (A) 9.8 m/s^2 ✓
 - (B) more than 9.8 m/s^2 ✗
 - (C) less than 9.8 m/s^2 ✗
 - (D) Cannot say unless the speed of throw is given. ✗

BigBench (Srivastava et al., 2023) includes ~ 200 tasks, some of which are multiple-choice and some require a short answer

Using humans in evaluation

Humans have been used in evaluation in text-generation tasks (including translation, summarization, dialogue)

Humans can **evaluate specific aspects** (e.g. fluency, accuracy) or make a **relative comparison** of two or more outputs

Chatbot Arena (Chiang et al., 2024) is a platform that uses crowdsourcing to compute a relative ranking of LLMs

Rank* (UB)	Model	Arena Elo	95% CI	Votes	Organization	License	Knowledge Cutoff
1	GPT-4-Turbo-2024-04-09	1259	+4/-3	35931	OpenAI	Proprietary	2023/12
2	GPT-4-1106-preview	1253	+2/-3	73547	OpenAI	Proprietary	2023/4
2	Claude 3 Opus	1251	+3/-3	80997	Anthropic	Proprietary	2023/8
2	Gemini 1.5 Pro API-0409-Preview	1250	+3/-3	39482	Google	Proprietary	2023/11
2	GPT-4-0125-preview	1247	+3/-2	67354	OpenAI	Proprietary	2023/12
6	Llama-3-70b-Instruct	1210	+3/-4	53404	Meta	Llama 3 Community	2023/12
6	Bard (Gemini Pro)	1209	+5/-6	12387	Google	Proprietary	Online

✂ Chatbot Arena: Benchmarking LLMs in the Wild

| [Blog](#) | [GitHub](#) | [Paper](#) | [Dataset](#) | [Twitter](#) | [Discord](#) |

📋 Rules

- Ask any question to two anonymous models (e.g., ChatGPT, Claude, Llama) and vote for the better one!
- You can continue chatting until you identify a winner.
- Vote won't be counted if model identity is revealed during conversation.

🏆 Arena Elo [Leaderboard](#)

We collect 300K+ human votes to compute an Elo-based LLM leaderboard. Find out who is the 🏆 LLM Champion!

🗨 Chat now!

Expand to see the descriptions of 36 models

Model A

i found a property i would like to buy. please draft a letter to the owners expressing my interest to purchase and convincing them of completing a private sale with me. list the benefits of private sale and how it will be done with our lawyers.

Subject: Expression of Interest in Your Property for a Private Sale

Dear [Owner's Name],

I hope this message finds you well. My name is [Your Name], and I am writing to express my sincere interest in purchasing your property located at [Property Address]. I have recently had the privilege of viewing your property and am captivated by its unique features and the potential it holds.

I would like to propose a private sale as a method of transaction. This proposal aims to create a beneficial situation for both parties involved, and I am confident that this process will be both smooth and efficient.

The advantages of a private sale are numerous. Firstly, it offers a more personal approach, as it allows for direct negotiation between the buyer and seller, resulting in potentially better terms for both parties. Secondly, a private sale can generally be completed more quickly than traditional sales, as there are fewer parties involved in the decision-making process. Thirdly, the

Model B

i found a property i would like to buy. please draft a letter to the owners expressing my interest to purchase and convincing them of completing a private sale with me. list the benefits of private sale and how it will be done with our lawyers.

Subject: Expression of Interest to Purchase Your Property - Private Sale Proposal

Dear [Owner's Name],

I hope this letter finds you well. My name is [Your Name], and I recently had the opportunity to view your property at [Property Address]. I am writing to express my genuine interest in purchasing your property and to propose the possibility of completing this sale privately, directly between us, with the support of our respective lawyers.

Firstly, I want to commend you on the care and upkeep of your property. It's clear that a great deal of effort has been invested into making it a desirable and welcoming space. After careful consideration and viewing several properties, I am convinced that your property meets all the criteria I am looking for in a home.

I understand that selling a property is a significant decision, and I believe that a private sale could offer considerable benefits to both of us. Here are a few reasons why I think a private sale would be advantageous:

👉 A is better

👉 B is better

👉 Tie


👉 Both are bad

Using an LLM as an automatic evaluator

Recently, “**LLM-as-a-judge**” methods have become popular

Zheng et al. (2023) claim that GPT-4 rankings largely agree with human rankings, but have some biases







AlpacaEval controls for length to avoid length bias in LLM evaluators (Dubois et al., 2024)

AlpacaEval  Leaderboard

An Automatic Evaluator for Instruction-following Language Models
Length-controlled (LC) win rates alleviate length biases of GPT-4, but it may favor models finetuned on its outputs.

Version: AlpacaEval AlpacaEval 2.0 Filter: Community Verified

Baseline: Davinci003 | Auto-annotator: GPT-4

Model Name	LC Win Rate	Win Rate
GPT-4 Preview (11/06) 	89.9%	97.7%
Mistral Medium 	91.5%	96.8%
GPT-4 	86.5%	95.3%
Mixtral 8x7B v0.1 	82.6%	94.8%
GPT-4 (03/14) 	85.3%	94.8%
Yi 34B Chat 	76.4%	94.1%

Main points: recap

LLM evaluation is designed to test their **general usability**

Benchmarks typically evaluate a wide range of capabilities

Evaluation formats range from simple answers to free text-generation

It is becoming more common to use **LLMs as evaluators**

References

- G. Bai, J. Liu, and X. Bu et al. 2024. [MT-Bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues](#). arXiv:2402.14762.
- W.-L. Chiang, L. Zheng, and Y. Sheng et al. 2024. [Chatbot arena: An open platform for evaluating LLMs by human preference](#). arXiv:2403.04132.
- A. Deshpande, V. Murahari, and T. Rajpurohit et al. 2023. [Toxicity in ChatGPT: Analyzing persona-assigned language models](#). In *EMNLP Findings*.
- Y. Dubois, B. Galambosi, P. Liang, and T. Hashimoto. 2024. [Length-controlled AlpacaEval: A simple way to debias automatic evaluators](#). arXiv:2404.04475.
- D. Hendrycks, C. Burns, and S. Basart et al. 2021. [Measuring massive multitask language understanding](#). In *ICLR*.
- F. Petroni, T. Rocktäschel, and S. Riedel et al. 2019. [Language models as knowledge bases?](#) In *EMNLP-IJCNLP*.
- A. Srivastava, A. Rastogi, and A. Rao et al. 2023. [Beyond the imitation game: Quantifying and extrapolating the capabilities of language models](#). *TMLR*.
- A. Wang, A. Singh, and J. Michael et al. 2019. [GLUE: A multi-task benchmark and analysis platform for natural language understanding](#). In *ICLR*.
- R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi. 2019. [HellaSwag: Can a machine really finish your sentence?](#) In *ACL*.
- L. Zheng, W.-L. Chiang, and Y. Sheng et al. 2023. [Judging LLM-as-a-judge with MT-Bench and Chatbot Arena](#). In *NeurIPS*.