

自然语言理解

WELCOME!

杨海钦

2025-2026-1学期

面试问题

- 什么是数据标注，为什么需要数据标注？
- 数据标注的流程是怎样的？如何控制标注质量？
- 什么是标注者间一致性？如果团队的一致性较低，可能的原因是什么？如何提升？

大纲

- 系统概览
- 数据处理与标注
 - 预训练技术
 - 大语言模型评测
 - 具体例子

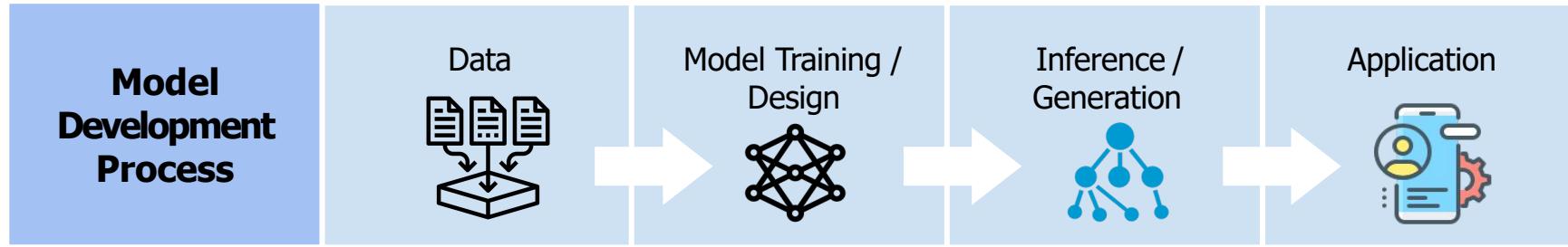
自然语言处理系统通用算法框架

- 创建一个函数将输入 X 映射到输出 Y , 其中 X 和/或 Y 涉及语言

- 任务

输入 X	输出 Y	任务
文本	连续文本	语言模型 Language modeling
文本	其他语言的文本	机器翻译 Machine Translation
文本	摘要	自动摘要 Automatic Summarization
文本	标签	文本分类 Text Classification
文本	语言结构	语言分析 Language Analysis
语音/ 文本	文本/ 语音	语音识别 ASR (Automatic Speech Recognition) /语音合成 TTS (Text-to-Speech)
图片/ 文本	文本/ 图片(视频)	图片描述 Image Captioning /文生图 T2I (文生视频 T2V)

自然语言处理系统开发流程



预测三步骤/改进五步骤

1. 特征提取：从文本中提取出决策所需的显著特征
 - 特征提取: $h = f(x)$
2. 分数计算：计算一个或多个可能性的分数
 - 分数计算:
 - 二分类: $s = w \cdot h$
 - 多分类: $s = Wh$
3. 决策函数：从几种可能性中选择一种
 - 决策函数:
 - $\hat{y} = \text{decide}(s/s)$
4. 精度计算 Accuracy Calculation
5. 错误分析 Error Analysis

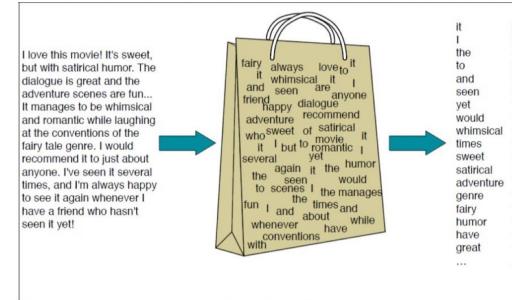
基于规则与基于学习(词袋Bag-of-Words, BOW)



```
In [1]: def extract_features(x: str) -> dict[str, float]:
    features = {}
    x_split = x.split(' ')
    good_words = ['love', 'good', 'nice', 'great', 'enjoy', 'enjoyed']
    bad_words = ['hate', 'bad', 'terrible', 'disappointing', 'sad', 'lost', 'angry']
    for x_word in x_split:
        if x_word in good_words:
            features['good_word_count'] = features.get('good_word_count', 0) + 1
        if x_word in bad_words:
            features['bad_word_count'] = features.get('bad_word_count', 0) + 1
    features['bias'] = 1
    return features
feature_weights = {'good_word_count': 1.0, 'bad_word_count': -1.0, 'bias': 0.5}
```

Train accuracy: 0.4345739700374532
Dev/test accuracy: 0.4214350590372389

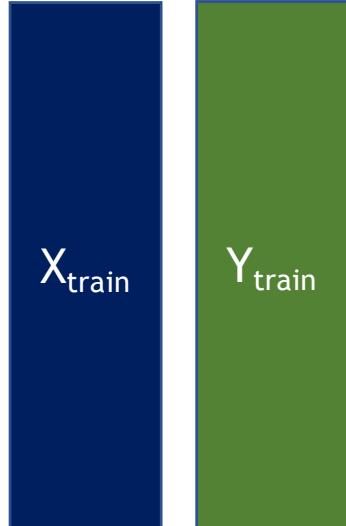
<https://github.com/hqyang/nlp-codes/tree/main/01-simpleclassifier>



```
feature_weights = {}
for x, y in data:
    # Make a prediction
    features = extract_features(x)
    predicted_y = run_classifier(features)
    # Update the weights if the prediction is wrong
    if predicted_y != y:
        for feature in features:
            feature_weights[feature] = (
                feature_weights.get(feature, 0) +
                y * features[feature]
            )
```

Train accuracy: 0.7332631086142322
Dev/test accuracy: 0.5676657584014533

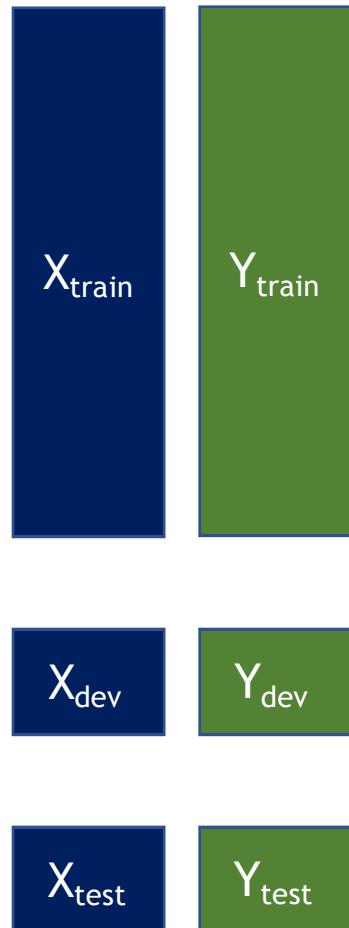
构建三组数据集



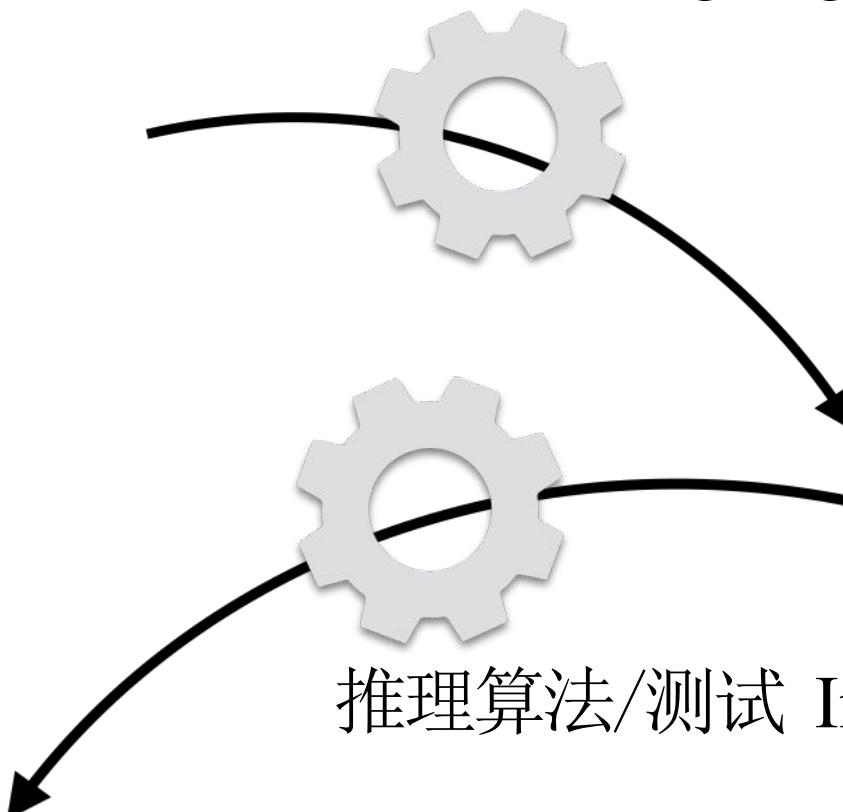
- **训练集(training set):** 通常数据集较大，在系统设计、创建和参数学习时使用
- **开发集/验证集(dev/validation set):** 用于测试不同设计决策(获得超参数)的较小数据集
- **测试集(test set):** 反映最终测试场景的数据集，用于评估模型性能



机器学习/Machine Learning



学习算法/训练 Learning Algorithm/Training



可学习的特征提取器 f
打分函数权值 w

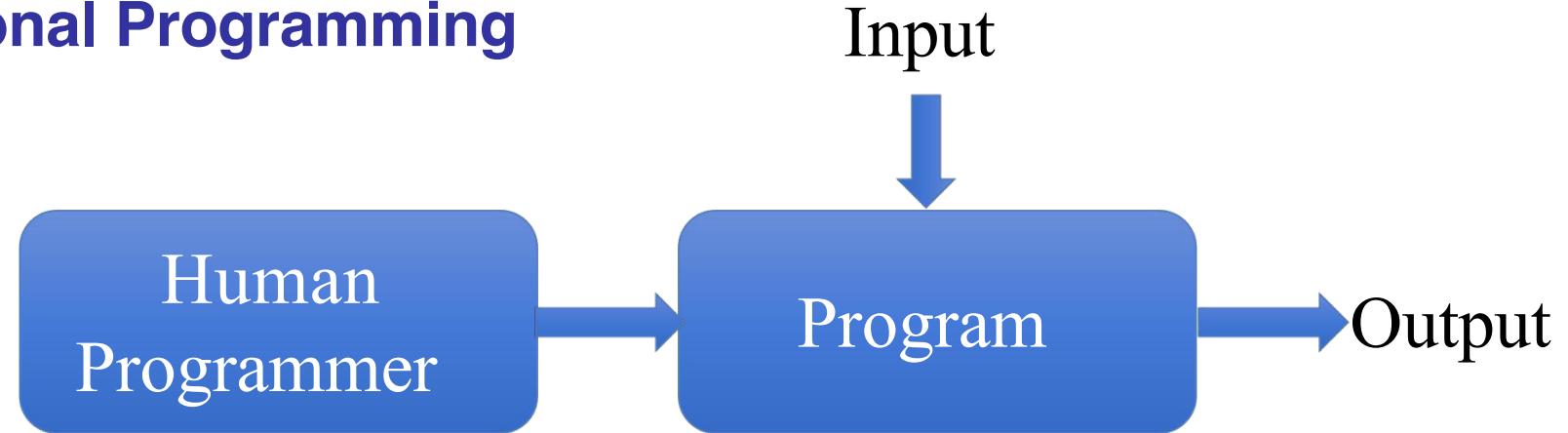
$$h = f(x)$$
$$s = w \cdot h$$

推理算法/测试 Inference Algorithm/Test

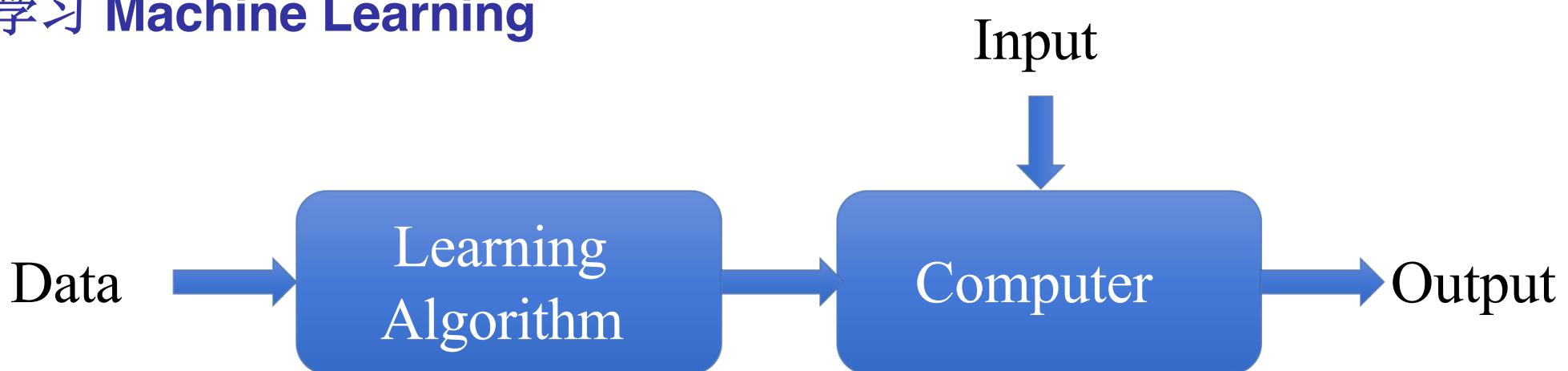
通过经验(数据)自动改进计算机算法 – Tom Mitchell

传统编程与机器学习

传统编程 **Traditional Programming**



机器学习 **Machine Learning**



神奇吗？

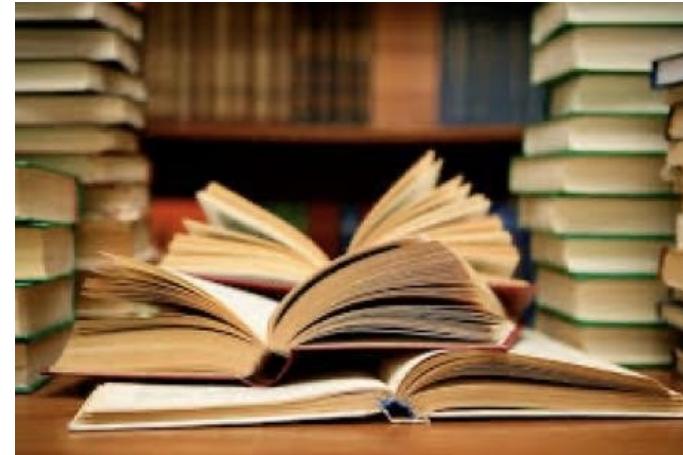
- No, more like gardening
- **Seeds** = Algorithms
- **Nutrients** = Data
- **Gardener** = You
- **Plants** = Programs



Credited from Pedro Domingos

高质量数据哪里来？

- 预训练数据
- 领域数据



预训练数据清洗

Model	Size (# Parameters)	Training Tokens
LaMDA (Thoppilan et al., 2022)	137 Billion	168 Billion
GPT-3 (Brown et al., 2020)	175 Billion	300 Billion
Jurassic (Lieber et al., 2021)	178 Billion	300 Billion
Gopher (Rae et al., 2021)	280 Billion	300 Billion
MT-NLG 530B (Smith et al., 2022)	530 Billion	270 Billion
<i>Chinchilla</i>	70 Billion	1.4 Trillion

Jordan Hoffman and et al. (2022). Training Compute-Optimal Large Language Models. <https://arxiv.org/pdf/2203.15556>

H. Touvron and et al. (2023) LLaMA: Open and Efficient Foundation Language Models. <https://arxiv.org/pdf/2302.13971>

	Disk Size	Documents	Sampling proportion	Epochs in 1.4T tokens
MassiveWeb	1.9 TB	604M	45% (48%)	1.24
Books	2.1 TB	4M	30% (27%)	0.75
C4	0.75 TB	361M	10% (10%)	0.77
News	2.7 TB	1.1B	10% (10%)	0.21
GitHub	3.1 TB	142M	4% (3%)	0.13
Wikipedia	0.001 TB	6M	1% (2%)	3.40

Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

C4: The Colossal Clean Crawled Corpus

预训练数据

- 低质内容
 - 社媒噪声
 - 拼写错误: Departmen → Department
 - 表情符误用: “ ” 表示愤怒而非悲伤
 - 情绪化表述: “你懂个屁!” “小编脑残”
 - 冗余信息:
 - 用户签名档: “关注我获取更多理财技巧”
 - 重复发帖
 - 广告: 代写标书
- 重复与模板化内容
- 格式混乱: $E=mc^2/E=mc2$
- 标签与无效代码
 - 未闭合标签: <div>内容</div>
- ...

C4: The Colossal Clean Crawled Corpus

- Source: Common Crawl (<https://commoncrawl.org/>)
- 用途: 常识推理、机器翻译、预训练数据、优化器测试集等
- 启发式清洗过程
 - 只保留以结束标点符号（即句号、感叹号、问号或结束引号）结尾的行
 - 丢弃任何少于3句话的页面，只保留包含至少5个单词的行
 - 删除包含脏词列表的页面(28种语言)
 - 删除任何含Javascript的一行
 - 占位符“lorem ipsum”(乱数假文): 删除所有出现“lorem ipsum”字样的页面
 - 删除了包含大括号的所有页面: 大括号“{”出现在编程语言中(如Javascript)，但不会出现在自然文本中
 - 删除引用标记: 由于一些被抓取的页面来自维基百科，并且有引用标记(例如[1], [xxx]等)
 - 删除样板政策通知行: 如含字符串“使用条款”，“隐私政策”，“cookie政策”，“使用cookie”，“使用cookie”或“使用cookie”的任何行
 - 删除重复数据: 丢弃数据集中出现超过一次的任何三句话跨度中的一个
 - 根据场景需要，使用[langdetect](#)选择相应语言文档

Colin Raffel et al. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. JMLR. <https://arxiv.org/abs/1910.10683>

DeepSeek

- Aggressive deduplication

Dumps Used	1	2	6	12	16	22	41	91
Deduplication Rate (%)	22.2	46.7	55.7	69.9	75.7	76.3	81.6	89.8

Table 1 | Deduplication ratios for various Common Crawl dumps.

- filtering
- remixing

DeepSeek (2024). DeepSeek LLM Scaling Open-Source Language Models with Longtermism.
<https://arxiv.org/pdf/2401.02954>

其他大模型清洗方案

- Qwen 千问
 - Qwen Technical Report, <https://arxiv.org/pdf/2309.16609>
 - Qwen2 Technical Report, <https://arxiv.org/pdf/2407.10671>
 - Qwen2.5 Technical Report, <https://arxiv.org/pdf/2412.15115>
- Gemini
 - Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities.
<https://arxiv.org/pdf/2507.0626>
 - [Qwen] Carefully curate high-quality training data through **sophisticated filtering** and **scoring mechanisms**, combined with strategic data mixture
 - Better: data filtering, math and code data, synthetic data, data mixture
 - expanding from the **7 trillion** tokens used in Qwen2 (Yang et al., 2024a) to **18 trillion tokens**

高质量数据哪里来？

- 预训练数据
- 领域数据

为什么需要构建领域数据？

- 应用驱动型研究：
 - 想做一个有用的系统，或者
 - 期望系统性能提升

大语言模型评测

- 评测概述
- 评测核心维度
- 主流评测方法
- 典型评测基准

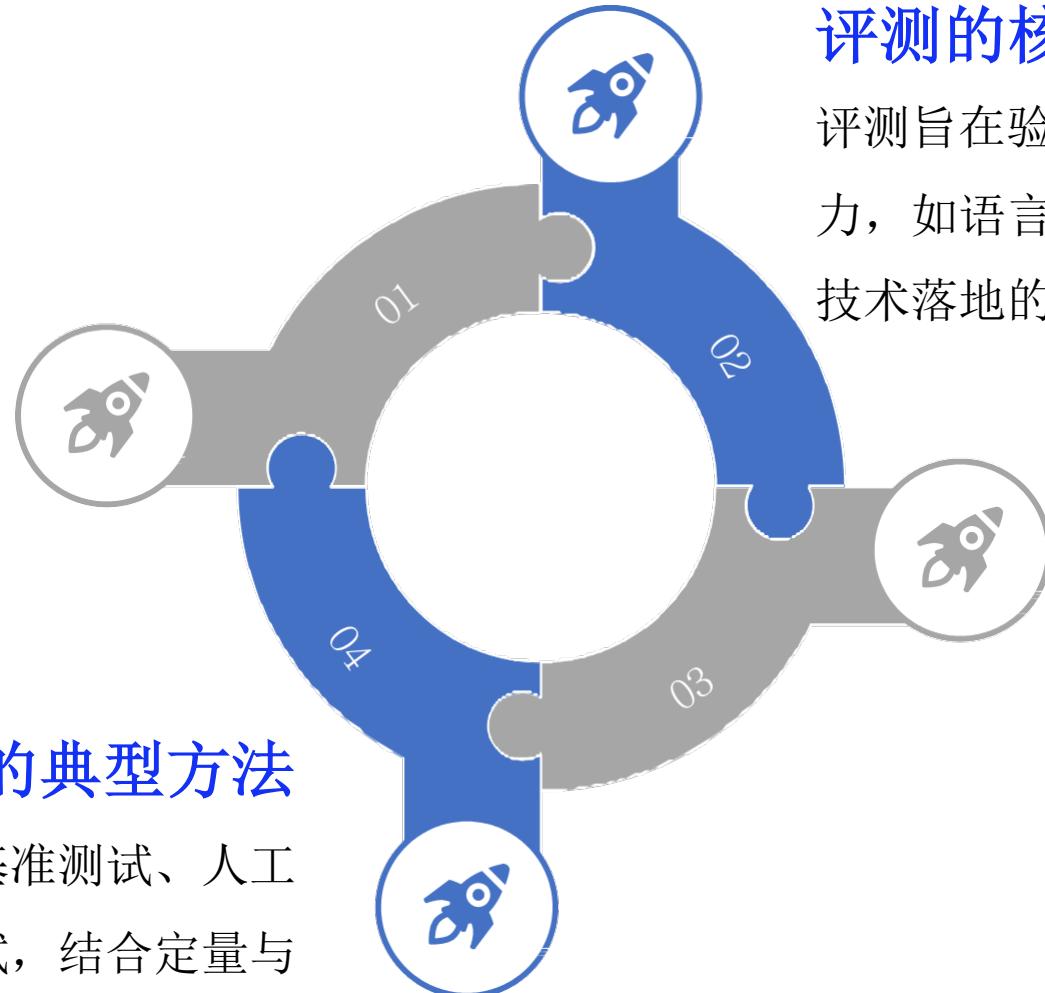
大语言模型评测

评测的基本概念

通过系统化方法衡量大模型性能的过程，包括准确性、效率等核心指标，为模型优化提供客观依据

评测的典型方法

评测方法包括基准测试、人工评估和对抗测试，结合定量与定性分析确保结果科学性



评测的核心目标

评测旨在验证大模型是否达到预期能力，如语言理解、生成质量等，确保技术落地的可靠性和实用性

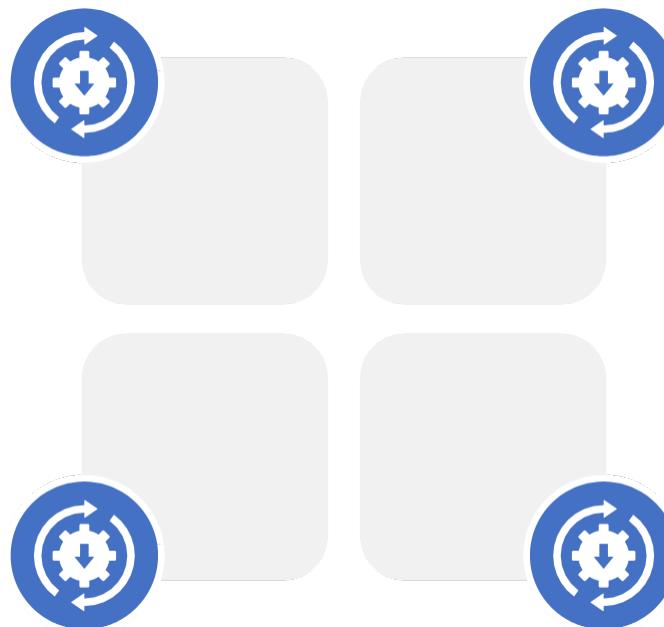
评测的关键维度

评测需覆盖多维度指标，包括任务完成度、泛化能力、鲁棒性等，全面反映模型综合性能

评测意义

衡量模型性能的标尺

评测通过量化指标客观反映大模型在各项任务中的表现，为技术迭代提供可比较的基准依据，是性能优化的核心参照



技术落地的质量保障

评测验证模型在真实场景中的可靠性，避免因技术缺陷导致应用风险，确保学术研究与产业应用的有效衔接

资源分配的决策依据

评测数据指导算力、数据等资源的精准投入，避免重复减少，提升大模型研发的效率与经济效益

评测挑战

- 1 评测标准的统一性 大模型评测缺乏统一标准，不同机构采用各异指标，导致结果难以横向比较，影响技术发展的客观评估
- 2 评测数据代表性不足 评测数据集覆盖场景有限，难以全面反映模型真实能力，可能掩盖模型在特定领域的缺陷或偏见
- 3 评测成本高昂 评测需消耗大量算力和时间资源，中小机构难以承担，形成技术垄断并阻碍创新公平性
- 4 动态演进的适配压力 模型计算迭代速度快，评测方法需持续更新以匹配创，滞后性可能导致评估结果失效新

大语言模型评测

- 评测概述
- 评测核心维度
- 主流评测方法
- 典型评测基准

能力维度



语言理解能力
评测大模型对语法、语义及上下文的理解深度，包括词义消歧、指代消解等任务，反映基础语言处理能力



逻辑推理能力
通过数学证明、常识推理等任务评估模型的多步推理能力，体现其从已知信息推导新结论的智能化水平



知识覆盖广度
测试模型对跨领域知识（如科学、历史）的掌握程度，揭示其训练数据多样性和知识泛化能力



生成质量与连贯性
衡量生成文本流畅度、逻辑性和创造性，包括长文本一致性及避免重读等关键指标

效率维度

计算资源效率

衡量大模型在训练和推理过程中对GPU/TPU等硬件资源的利用率，优化计算效率可显著降低能耗与成本

内存占用优化

评估模型运行时显存与内存的消耗量，通过量化、剪枝等技术减少内存压力，提升

部署灵活性



时间效率评估

分析模型完成特定任务所需的时间周期，包括训练收敛速度和推理响应延迟，直接影响实际应用可行性

能源能耗指标

量化模型训练与推理的电力损耗，绿色AI趋势下需平衡性能与碳排放，推动可持续发展

安全维度

大模型安全评测概述

大模型安全评测在评估模型在对抗攻击、隐私泄露等场景下的鲁棒性，是确保AI系统可靠性的关键环节

隐私数据保护机制

评估模型训练与推理过程中用户数据的脱敏处理、加密传输等隐私保护技术的完备性



对抗攻击防御能力

测试模型对对抗样本的识别与抵抗能力，包括文本对抗、图像对抗等攻击方式的防御效果评估

内容安全过滤性能

检测模型生成内容是否包含暴力、偏见等有害信息，验证安全过滤策略的有效性

大语言模型评测

- 评测概述
- 评测核心维度
- 主流评测方法
- 典型评测基准

人工评测



评测的定义与意义

人工评测指通过人类专家或标注员对大模型输出进行主观评估，是验证模型性能不可替代的核心手段，尤其擅长评估生成质量



评测维度设计原则

需设计多维度的评测指标体系，包括流畅性、相关性、事实准确性等，每个维度需明确定义评分标准以保证结果客观性



评测任务类型

涵盖开放生成、问答、摘要等任务类型，不同任务需定制化评测方案，例如对话系统需评估连贯性和意图理解能力



评测流程标准化

包含数据采样、标注培训、双盲评测等环节，需建立严格流程控制机制以降低主观偏差，确保结果可复现

自动评测

基本概念

利用算法和计算模型对系统性能进行量化评估的过程，无需人工干预即可完成高效客观的

测试



自动评测的核心技术

自动评测依赖自然语言处理、机器学习等技术，通过预设指标和标准对模型输出进行自动化分析与评分



自动评测的优势

自动评测具有高效性、可重复性和客观性，能够快速处理大量数据，显著降低人工评估的成本和时间



自动评测的常见指标

自动评测常用指标包括准确率、召回率、F1、BLEU、ROUGE等，从不同角度量化模型的性能表现

混合评测

混合评测的定义与意义

结合多种评估方法，全面衡量模型性能，既能检验单一指标，又能反映综合能力，提升评测结果的可靠性

混合评测的典型应用场景

评测方法包括基准测试、人工评估和对抗测试，结合定量与定性分析确保结果科学性



混合评测的核心方法

混合评测通常融合自动化测试与人工评估，兼顾效率与准确性，同时引入多维度指标，确保评估结果客观全面

混合评测的挑战与解决方案

混合评测面临成本高、标准不统一等问题，需通过优化流程、制定规范来平衡效率与质量，提升可操作性

大语言模型评测

- 评测概述
- 评测核心维度
- 主流评测方法
- 典型评测基准

人工智能评估的发展

- 传统时期(1990年代至2000年代): 小规模任务特定基准
 - MNIST(1998): 用于图像分类的手写数字识别
 - Penn Treebank(1993): NLP中句法分析的早期基准
 - Switchboard(1992): 语音识别评估数据集
- 深度学习时代(2010年代): 大规模特定任务基准
 - ImageNet(2010): 深度CNN革新了图像识别
 - COCO(2014): 复杂目标检测、分割和字幕
 - SQuAD(2016) 和SQuAD 2.0(2018): 机器阅读理解和QA
 - GLUE(2018) 和SuperGLUE(2019): 多个NLU任务测试任务
- 基础模型(Foundation model)时代(2020年代):
 - MMLU(2020): 测试不同领域的推理和事实知识
 - Big-bench(2021): 推理、世界知识和创造力
 - MMMU(2023): 多主题的复杂推理多模态问题

大语言模型典型评测基准

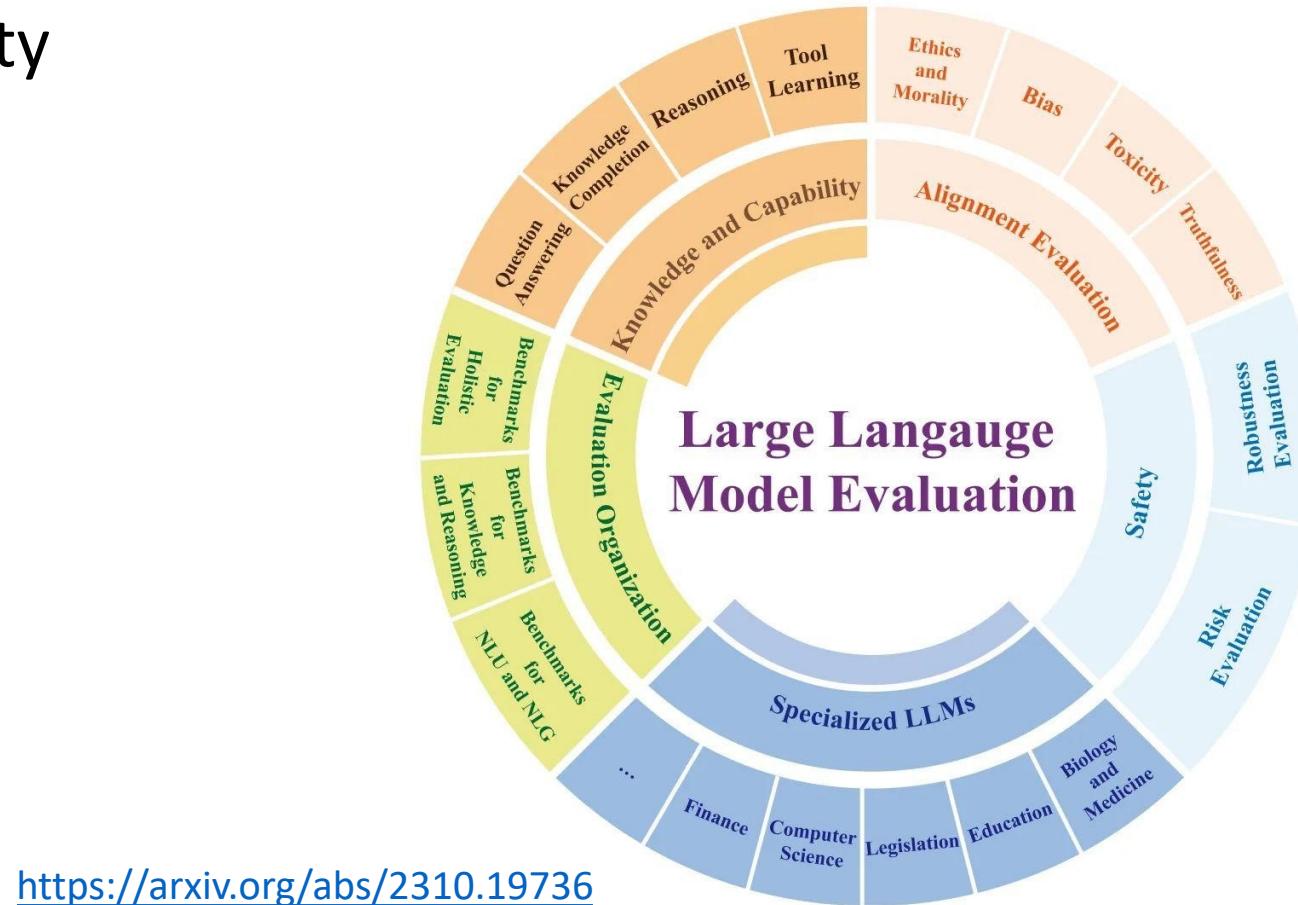
基准名称	发布时间	发布人/机构	规模	评价方面	语种数					
DecaNLP	2018 年	McCann 等	10 个任务	NLU	1					
GLUE	2018 年	纽约大学等	9 个任务	NLU	1					
SuperGLUE	2019 年	纽约大学等	8 个任务	NLU	1					
XTREME	2020 年	卡耐基梅隆大学等	9 个任务	NLU	40					
XGLUE	2020 年	微软	11 个任务	NLU	Chatbot Arena	2023 年	伯克利大学等	—	与人类偏好对齐程度	—
CLUE	2020 年	CLUE 团队	9 个任务	NLU	SuperCLUE-Open	2023 年	CLUE 团队	1.2k 个样例	与人类偏好对齐程度	1
GLGE	2020 年	四川大学等	4 个任务	NLG	C-EVAL	2023 年	上海交通大学等	13948 个样例	知识运用与推理能力	1
GEM	2021 年	卡耐基梅隆大学等	11 个数据集	NLG	SAFETYPROMPTS	2023 年	清华大学等	100k 个样例	安全性	1
CUGE	2021 年	清华大学等	21 个数据集	NLU、NLG	LLMEVAL	2023 年	复旦大学等	933 个样例	正确性、流畅性、信息量、逻辑性、无害性等	1
EleutherAI LM Harness	2021 年	EleutherAI	200 余个数据集	NLU、NLG	TriviaQA	2017 年	华盛顿大学等	95k 个样例	知识运用能力	1
BIG-bench	2022 年	谷歌	200 余个数据集	逻辑能力、数学能力、代码理 伦理道德	OpenBookQA	2018 年	艾伦人工智能 研究所等	约 6k 个样例	知识运用能力	1
MT-Bench	2023 年	伯克利大学等	80 个样例	与人类偏好对齐程度	GSM8k	2021 年	OpenAI 等	8.5k 个样例	数学推理能力	1
					HaluEval	2023 年	人民大学等	35k 个样例	检测幻觉能力	1
					MMLU	2021 年	伯克利大学等	57 个数据集	知识运用与问题解决能力	1
					HellaSwag	2019 年	华盛顿大学	70k 个样例	常识推理能力	1
					HumanEval	2021 年	OpenAI 等	164 个样例	代码生成能力	—
					DROP	2019 年	加利福尼亚大学等	96k 个样例	阅读理解和数值推理能力	1

Source: [53AI](#)



大语言模型典型评测基准

- Knowledge and Capability
- Alignment Evaluation
- Safety
- Specialized LLMs
- Evaluation Organization



General Language Understanding Evaluation (GLUE)

Corpus	Train	Test	Task	Metrics	Domain
Single-Sentence Tasks					
CoLA	8.5k	1k	acceptability	Matthews corr.	misc.
SST-2	67k	1.8k	sentiment	acc.	movie reviews
Similarity and Paraphrase Tasks					
MRPC	3.7k	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	391k	paraphrase	acc./F1	social QA questions
Inference Tasks					
MNLI	393k	20k	NLI	matched acc./mismatched acc.	misc.
QNLI	105k	5.4k	QA/NLI	acc.	Wikipedia
RTE	2.5k	3k	NLI	acc.	news, Wikipedia
WNLI	634	146	coreference/NLI	acc.	fiction books

Table 1: Task descriptions and statistics. All tasks are single sentence or sentence pair classification, except STS-B, which is a regression task. MNLI has three classes; all other classification tasks have two. Test sets shown in bold use labels that have never been made public in any form.

<https://arxiv.org/pdf/1804.07461>

SuperGLUE

Table 1: The tasks included in SuperGLUE. *WSD* stands for word sense disambiguation, *NLI* is natural language inference, *coref.* is coreference resolution, and *QA* is question answering. For MultiRC, we list the number of total answers for 456/83/166 train/dev/test questions.

Corpus	Train	Dev	Test	Task	Metrics	Text Sources
BoolQ	9427	3270	3245	QA	acc.	Google queries, Wikipedia
CB	250	57	250	NLI	acc./F1	various
COPA	400	100	500	QA	acc.	blogs, photography encyclopedia
MultiRC	5100	953	1800	QA	F1 _a /EM	various
ReCoRD	101k	10k	10k	QA	F1/EM	news (CNN, Daily Mail)
RTE	2500	278	300	NLI	acc.	news, Wikipedia
WiC	6000	638	1400	WSD	acc.	WordNet, VerbNet, Wiktionary
WSC	554	104	146	coref.	acc.	fiction books

<https://arxiv.org/pdf/1905.00537>

Massive Multitask Language Understanding (MMLU)

Few Shot Prompt and Predicted Answer

The following are multiple choice questions about high school mathematics.

How many numbers are in the list 25, 26, ..., 100?

- (A) 75 (B) 76 (C) 22 (D) 23

Answer: B

Compute $i + i^2 + i^3 + \dots + i^{258} + i^{259}$.

- (A) -1 (B) 1 (C) i (D) $-i$

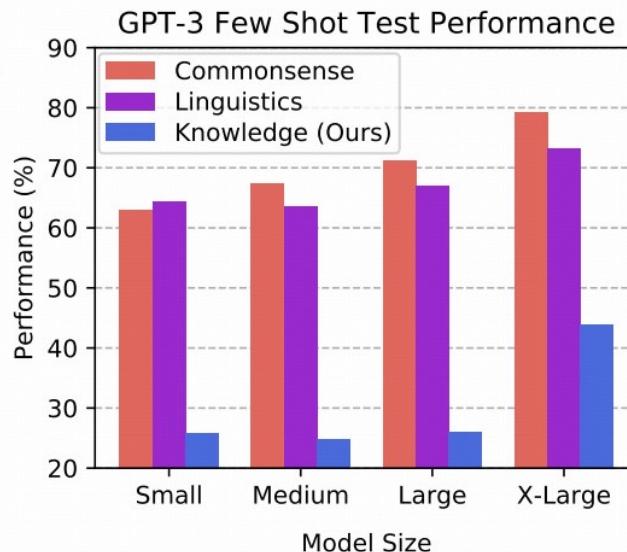
Answer: A

If 4 daps = 7 yaps, and 5 yaps = 3 baps,
how many daps equal 42 baps?

- (A) 28 (B) 21 (C) 40 (D) 30

Answer: C

(a) An example of few-shot learning and inference using GPT-3. The blue underlined bold text is the auto-completed response from GPT-3, while the preceding text is the user-inputted prompt. In this 2-shot learning example, there are two instruction examples and one initially incomplete example. On average, GPT-3 has low accuracy on high school mathematics questions.



(b) Performance on a commonsense benchmark (HellaSwag), a linguistic understanding benchmark (SuperGLUE), and the massive multitask test. On previous benchmarks, smaller models start well above random chance levels and exhibit more continuous improvements with model size increases, but on our test, GPT-3 moves beyond random chance with the largest model.

<https://arxiv.org/pdf/2009.03300>

Massive Multitask Language Understanding (MMLU)

Task	Tested Concepts	Supercategory			
Abstract Algebra	Groups, rings, fields, vector spaces, ...	STEM	High School Statistics	Random variables, sampling distributions, chi-square tests, ...	STEM
Anatomy	Central nervous system, circulatory system, ...	STEM	High School US History	Civil War, The Great Depression, The Great Society, ...	Humanities
Astronomy	Solar system, galaxies, asteroids, ...	STEM	High School World History	Ottoman empire, economic imperialism, World War I, ...	Humanities
Business Ethics	Corporate responsibility, stakeholders, regulation, ...	Other	Human Aging	Senescence, dementia, longevity, personality changes, ...	Other
Clinical Knowledge	Spot diagnosis, joints, abdominal examination, ...	Other	Human Sexuality	Pregnancy, sexual differentiation, sexual orientation, ...	Social Sciences
College Biology	Cellular structure, molecular biology, ecology, ...	STEM	International Law	Human rights, sovereignty, law of the sea, use of force, ...	Humanities
College Chemistry	Analytical, organic, inorganic, physical, ...	STEM	Jurisprudence	Natural law, classical legal positivism, legal realism, ...	Humanities
College Computer Science	Algorithms, systems, graphs, recursion, ...	STEM	Logical Fallacies	No true Scotsman, base rate fallacy, composition fallacy, ...	Humanities
College Mathematics	Differential equations, real analysis, combinatorics, ...	STEM	Machine Learning	SVMs, VC dimension, deep learning architectures, ...	STEM
College Medicine	Introductory biochemistry, sociology, reasoning, ...	Other	Management	Organizing, communication, organizational structure, ...	Other
College Physics	Electromagnetism, thermodynamics, special relativity, ...	STEM	Marketing	Segmentation, pricing, market research, ...	Other
Computer Security	Cryptography, malware, side channels, fuzzing, ...	STEM	Medical Genetics	Genes and cancer, common chromosome disorders, ...	Other
Conceptual Physics	Newton's laws, rotational motion, gravity, sound, ...	STEM	Miscellaneous	Agriculture, Fermi estimation, pop culture, ...	Other
Econometrics	Volatility, long-run relationships, forecasting, ...	Social Sciences	Moral Disputes	Freedom of speech, addiction, the death penalty, ...	Humanities
Electrical Engineering	Circuits, power systems, electrical drives, ...	STEM	Moral Scenarios	Detecting physical violence, stealing, externalities, ...	Humanities
Elementary Mathematics	Word problems, multiplication, remainders, rounding, ...	STEM	Nutrition	Metabolism, water-soluble vitamins, diabetes, ...	Other
Formal Logic	Propositions, predicate logic, first-order logic, ...	Humanities	Philosophy	Skepticism, phrenology, skepticism, Singer's Drowning Child, ...	Humanities
Global Facts	Extreme poverty, literacy rates, life expectancy, ...	Other	Prehistory	Neanderthals, Mesoamerica, extinction, stone tools, ...	Humanities
High School Biology	Natural selection, heredity, cell cycle, Krebs cycle, ...	STEM	Professional Accounting	Auditing, reporting, regulation, valuation, ...	Other
High School Chemistry	Chemical reactions, ions, acids and bases, ...	STEM	Professional Law	Torts, criminal law, contracts, property, evidence, ...	Humanities
High School Computer Science	Arrays, conditionals, iteration, inheritance, ...	STEM	Professional Medicine	Diagnosis, pharmacotherapy, disease prevention, ...	Other
High School European History	Renaissance, reformation, industrialization, ...	Humanities	Professional Psychology	Diagnosis, biology and behavior, lifespan development, ...	Social Sciences
High School Geography	Population migration, rural land-use, urban processes, ...	Social Sciences	Public Relations	Media theory, crisis management, intelligence gathering, ...	Social Sciences
High School Gov't and Politics	Branches of government, civil liberties, political ideologies, ...	Social Sciences	Security Studies	Environmental security, terrorism, weapons of mass destruction, ...	Social Sciences
High School Macroeconomics	Economic indicators, national income, international trade, ...	Social Sciences	Sociology	Socialization, cities and community, inequality and wealth, ...	Social Sciences
High School Mathematics	Pre-algebra, algebra, trigonometry, calculus, ...	STEM	US Foreign Policy	Soft power, Cold War foreign policy, isolationism, ...	Social Sciences
High School Microeconomics	Supply and demand, imperfect competition, market failure, ...	Social Sciences	Virology	Epidemiology, coronaviruses, retroviruses, herpesviruses, ...	Other
High School Physics	Kinematics, energy, torque, fluid pressure, ...	STEM	World Religions	Judaism, Christianity, Islam, Buddhism, Jainism, ...	Humanities
High School Psychology	Behavior, personality, emotions, learning, ...	Social Sciences			

Summary of all 57 tasks

<https://arxiv.org/pdf/2009.03300>

MMLU - 追踪大模型性能进展

Closed-source vs. open-weight models

Llama 3.1 405B closes the gap with closed-source models for the first time in history.

@maximelabonne



<https://x.com/maximelabonne/status/1816416043511808259>

Big Bench

Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models

Alphabetic author list:^{*}

 BIG-bench

The Beyond the Imitation Game Benchmark (BIG-bench) is a collaborative benchmark intended to probe large language models and extrapolate their future capabilities. The more than 200 tasks included in BIG-bench are summarized by keyword [here](#), and by task name [here](#). A paper introducing the benchmark, including evaluation results on large language models, is currently under review, and is available as a [preprint](#).

The benchmark organizers can be contacted at bigbench@googlegroups.com.

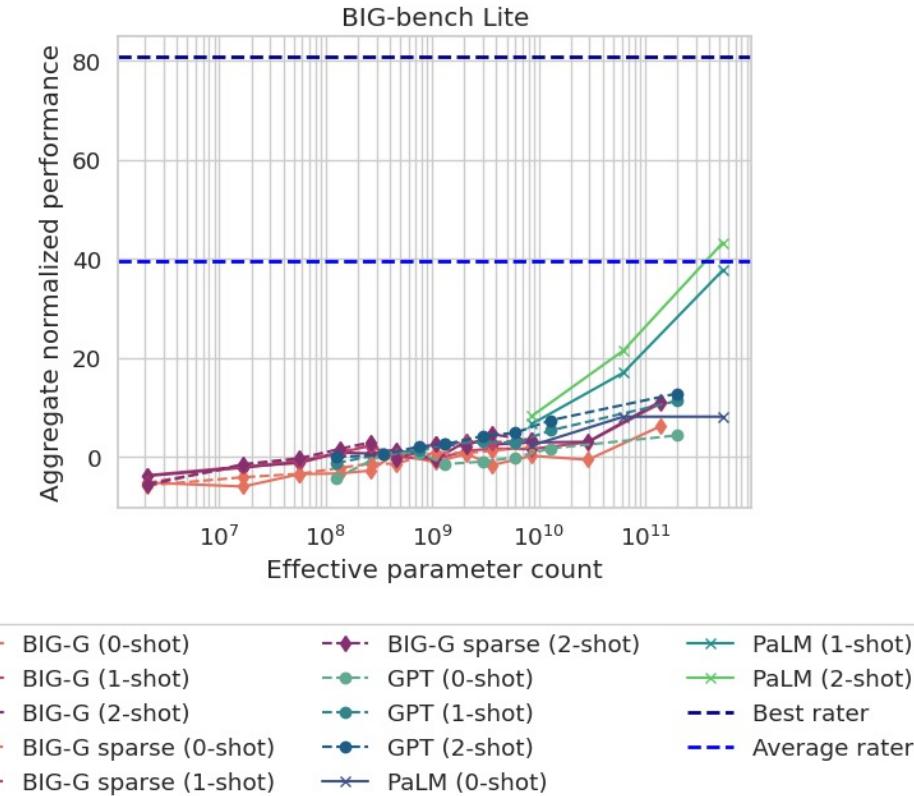
Table of contents

- [BIG-bench Lite leaderboard](#)
 - [Quick start](#)
 - [Installation](#)
 - [How do I create a task?](#)
 - [Creating a programmatic task](#)
 - [Submitting a model evaluation](#)
 - [Frequently asked questions](#)
 - [Alan Turing sitting on a bench](#)

For more details about the benchmark, see our [detailed instructions](#).



<https://arxiv.org/pdf/2206.04615>



Big Bench Hard (BBH)

# Tasks	Criteria
<u>209</u>	<u>All BIG-Bench tasks</u>
187	- After filtering out tasks with more than three subtasks
130	- After filtering out tasks with fewer than 103 examples (3 for few-shot, 100 for evaluation)
85	- After filtering out tasks without human-rater baselines
78	- After filtering out tasks that do not use multiple-choice or exact match as the evaluation metric
<u>78</u>	<u>Clean multiple-choice or exact match tasks</u>
36	- After filtering out tasks in which the best reported model beats average reported human-rater score
23	- After filtering out extremely difficult tasks that are outside the scope of this work
23	Remaining tasks = BIG-Bench Hard (BBH)

Table 1: Filtering criteria to used to create the BIG-Bench Hard (BBH) subset. Exact names of the BIG-Bench tasks filtered out by each criteria are shown in Appendix D.

<https://arxiv.org/pdf/2210.09261>

Big Bench Hard (BBH)

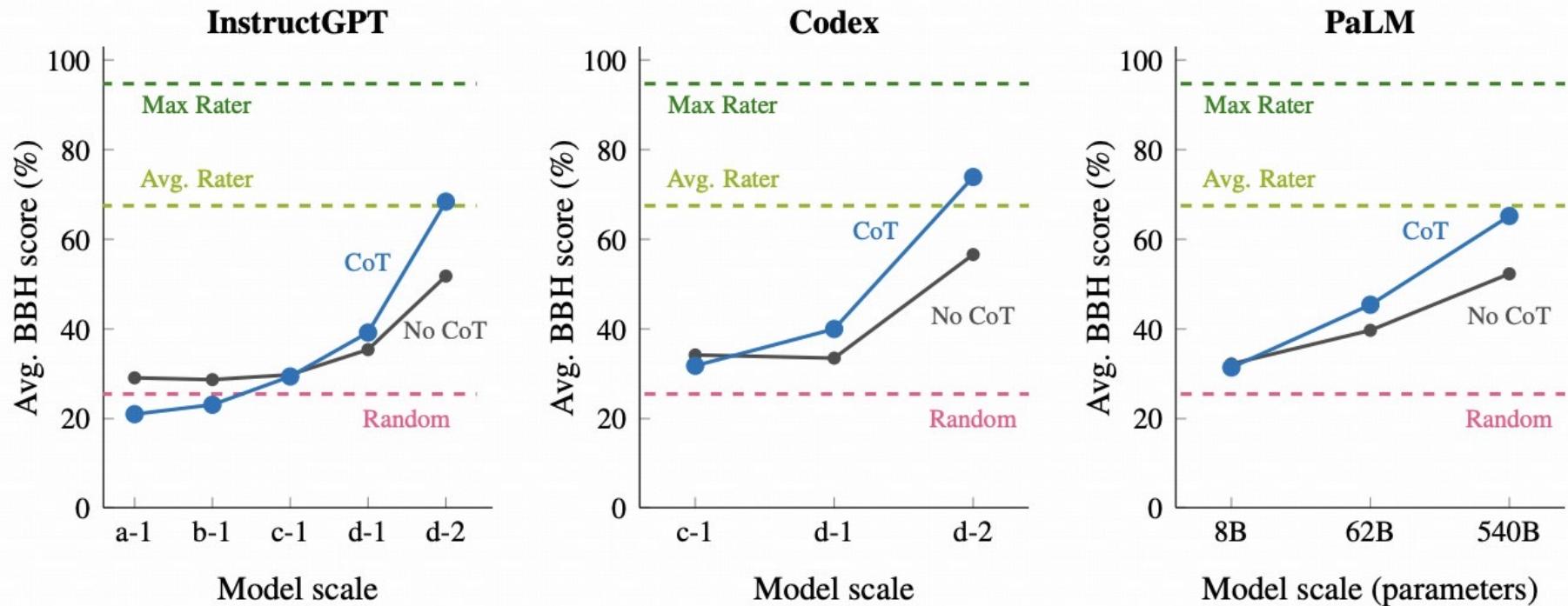


Figure 4: Scaling behavior of chain-of-thought (CoT) prompting on BIG-Bench Hard (BBH; 23 task unweighted average). InstructGPT models are the following: a-1 (text-ada-001), b-1 (text-babbage-001), c-1 (text-curie-001), d-1 (text-davinci-001), and d-2 (text-davinci-002). Codex models are the following: c-1 (code-cushman-001), d-1 (code-davinci-002), and d-2 (code-davinci-002).

<https://arxiv.org/pdf/2210.09261>

HELM: Holistic Evaluation of Language Models

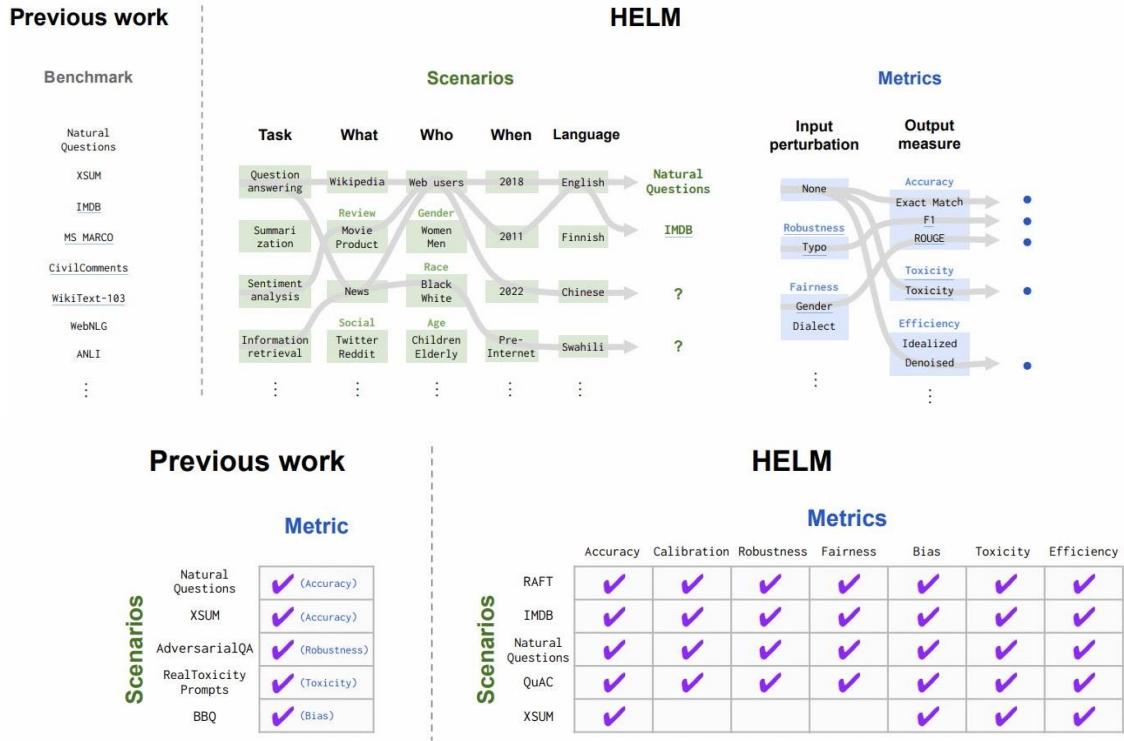


Figure 3: **Many metrics for each use case.** In comparison to most prior benchmarks of language technologies, which primarily center accuracy and often relegate other desiderata to their own bespoke datasets (if at all), in HELM we take a multi-metric approach. This foregrounds metrics beyond accuracy and allows one to study the tradeoffs between the metrics.

- 自顶向下
- 统一评估框架/平台
- 7大评估指标
 - Accuracy
 - Calibration
 - Robustness
 - Fairness
 - Bias
 - Toxicity
 - Efficiency
- 数十核心场景：问答、推理、摘要、对话等

<https://arxiv.org/pdf/2211.09110>

Chatbot Arena: Evaluation by Human Preference

The screenshot shows a web-based platform for evaluating Large Language Models (LLMs) through human preference. At the top, there's a "Chat now!" button and a search bar with the placeholder "Expand to see the descriptions of 72 models". Below the search bar, two models are selected: "Model A" and "Model B".

Model A (claude-3-5-sonnet-20241022) and Model B (gemin-1.5-pro-002) are shown in separate panels. Each panel contains a "How to evaluate LLMs?" section with detailed information and a list of metrics.

A modal window titled "Info" appears on the right, containing a message: "Thanks for voting! Your vote shapes the leaderboard, please vote RESPONSIBLY." This message is accompanied by a small icon of a person with a speech bubble.

At the bottom of the interface, there are several buttons: "A is better", "B is better", "Tie", "Both are bad", "Send", "Random Image", "New Round", "Regenerate", and "Share". A note at the bottom says: "Press 'New Round' to start over. (Note: Your vote shapes the leaderboard, please vote RESPONSIBLY!)"

The URL <https://lmarena.ai/?mode=side-by-side> is displayed at the bottom of the screenshot.

Elo Rating System

The [Elo rating system](#) is a method for calculating the relative skill levels of players, which has been widely adopted in competitive games and sports. The difference in the ratings between two players serves as a predictor of the outcome of a match. The Elo rating system works well for our case because we have multiple models and we run pairwise battles between them.

If player A has a rating of R_A and player B a rating of R_B , the exact formula (using the logistic curve with base 10) for the probability of player A winning is

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}}.$$

The ratings of players can be linearly updated after each battle. Suppose player A (with Rating R_A) was expected to score E_A points but actually scored S_A points. The formula for updating that player's rating is

$$R'_A = R_A + K \cdot (S_A - E_A).$$

<https://blog.lmarena.ai/blog/2023/arena/>

In-class Practice I

- 在Chatbot Arena的大模型双盲对战中，采用Elo评分系统计算模型相对技能水平。假设：模型A当前Elo分数为1400，模型B当前Elo的分数为1500，Elo更新系数($K=32$)
 1. 计算模型A击败模型B的预期概率(E_A)；
 2. 假设A胜实际得分为 $S_A=1$ ，失败则 $S_A=0$ ，平局则 $S_A=0.5$ ，计算模型A在上述情况下更新后的Elo分数(R_A')
 3. 我们知道模型A的得分是B的失分，分析上述分数，可以得出什么结论？

Chatbot Arena Leaderboard

The screenshot shows the homepage of the Chatbot Arena Leaderboard. At the top, there's a navigation bar with links like Overview, Text, WebDev, Vision, etc., and a "Start Voting" button. Below the navigation is a section titled "Leaderboard Overview" with a brief description. The main content area features two leaderboards: "Text" and "WebDev". Each leaderboard has columns for Rank (UB), Model, Score, and Votes. The "Text" leaderboard shows the following data:

Rank (UB)	Model	Score	Votes
1	gemini-2.5-pro	1456	46,291
1	claude-opus-4-1-20250805-thi...	1449	14,843
2	o3-2025-04-16	1441	46,083
2	chatgpt-4o-latest-20250326	1441	35,807
2	gpt-4.5-preview-2025-02-27	1441	14,644
2	gpt-5-high	1440	17,636
2	claude-opus-4-1-20250805	1438	23,612
3	qwen3-max-preview	1430	11,851

The "WebDev" leaderboard shows the following data:

Rank (UB)	Model	Score	Votes
1	GPT-5 (high)	1483	4,330
1	Claude Opus 4.1 thinking-16k...	1476	2,303
3	Claude Opus 4.1 (20250805)	1448	2,639
4	Gemini-2.5-Pro	1405	7,854
4	DeepSeek-R1-0528	1393	4,800
5	Claude Opus 4 (20250514)	1383	9,238
5	GLM-4.5	1371	2,042
6	Qwen3-Coder	1367	7,184

<https://lmarena.ai/leaderboard>

LLM-as-Judge: MT-Bench

Question: If the FED buys bonds in the secondary market
(A) the money supply will increase. (B) the money supply will decrease. (C) the money supply will not be affected. (D) the discount rate would be affected. Output the answer only.

Assistant A:

The Federal Reserve buys bonds in the secondary market to increase the money supply.

Assistant B:

(A) the money supply will increase.

Follow-up Question: How does it affect my daily life? Give 3 examples.

Assistant A:

The Federal Reserve buys bonds in the secondary market to increase the money supply. This affects your daily life by:
1. Increasing the money supply.
2. Increasing the money supply and lowering the discount rate.
3. Increasing the money supply and lowering the discount rate and increasing the money supply.

Assistant B:

When the Federal Reserve buys bonds in the secondary market, it can have several effects on the economy that can impact your daily life. Here are three examples:

1. Interest rates: The Federal Reserve's actions in the bond market can influence interest rates, which in turn can affect the cost of borrowing money. For example, if the Federal Reserve buys bonds and drives up demand for them, interest rates on things like mortgages, car loans, and credit cards may go up, making it more expensive to borrow money.
2. Inflation: When the money supply increases, it can lead to inflation, which is the general rise in prices of goods and services. This can impact your daily life by making it more expensive to buy things like groceries, gas, and other necessities.
3. Employment: The Federal Reserve's actions in the bond market can also have an impact on employment. For example, if the Federal Reserve's actions lead to economic growth and increased demand for goods and services, it can create more job opportunities and boost employment.

GPT-4 Judgment:

Assistant A provided an incorrect response to the user's question about how the Federal Reserve buying bonds in the secondary market affects daily life. The answer given is repetitive and lacks clear examples of how the action impacts daily life.

On the other hand, Assistant B provided a relevant and accurate response to the user's question about the Federal Reserve buying bonds. The answer includes three clear examples of how the action impacts daily life, such as interest rates, inflation, and employment.

Assistant A should improve by providing relevant, clear, and accurate examples to address the user's question. Assistant B's response is more helpful, relevant, and detailed.

<https://arxiv.org/pdf/2306.05685>

LLM-as-Judge：利弊（Pros & Cons）

- **规模化和速度**: 大模型可以快速评估大量的响应或提交
- **成本效率**: 使用大模型比依赖多名人工评判成本更低，尤其是在大规模评估时
- **灵活性**: 大模型可以根据特定的标准或任务进行微调或调整，使他们能够根据不同的标准进行判断
- **误解的风险**: 如果措辞不寻常，大模型可能会误解输入。
- **易受操纵**: 大模型可能容易受到对抗性提示或“操纵”技术的影响（例如，越狱）。
- **对细微差别的理解有限**: 大模型可能很难理解评估中的主观、细微或依赖于上下文的因素

大纲

- 系统概览
- 数据处理与标注
 - 预训练技术
 - 大语言模型评测
 - 具体例子

General Language Understanding Evaluation (GLUE)

Corpus	Train	Test	Task	Metrics	Domain
Single-Sentence Tasks					
CoLA	8.5k	1k	acceptability	Matthews corr.	misc.
SST-2	67k	1.8k	sentiment	acc.	movie reviews
Similarity and Paraphrase Tasks					
MRPC	3.7k	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	391k	paraphrase	acc./F1	social QA questions
Inference Tasks					
MNLI	393k	20k	NLI	matched acc./mismatched acc.	misc.
QNLI	105k	5.4k	QA/NLI	acc.	Wikipedia
RTE	2.5k	3k	NLI	acc.	news, Wikipedia
WNLI	634	146	coreference/NLI	acc.	fiction books

Table 1: Task descriptions and statistics. All tasks are single sentence or sentence pair classification, except STS-B, which is a regression task. MNLI has three classes; all other classification tasks have two. Test sets shown in bold use labels that have never been made public in any form.

<https://arxiv.org/pdf/1804.07461>

情感分析

Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, July 2002, pp. 79-86.
Association for Computational Linguistics.

Thumbs up? Sentiment Classification using Machine Learning Techniques

Bo Pang and **Lillian Lee**
Department of Computer Science
Cornell University
Ithaca, NY 14853 USA
{pabo,llee}@cs.cornell.edu

Shivakumar Vaithyanathan
IBM Almaden Research Center
650 Harry Rd.
San Jose, CA 95120 USA
shiv@almaden.ibm.com

<https://aclanthology.org/W02-1011.pdf>

- Bo Pang, Lillian Lee, Shivakumar Vaithyanathan: Thumbs up? Sentiment Classification using Machine Learning Techniques. EMNLP 2002: 79-86
- Bo Pang, Lillian Lee: A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. ACL 2004: 271-278
- Bo Pang, Lillian Lee: Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. ACL 2005: 115-124

情感分析

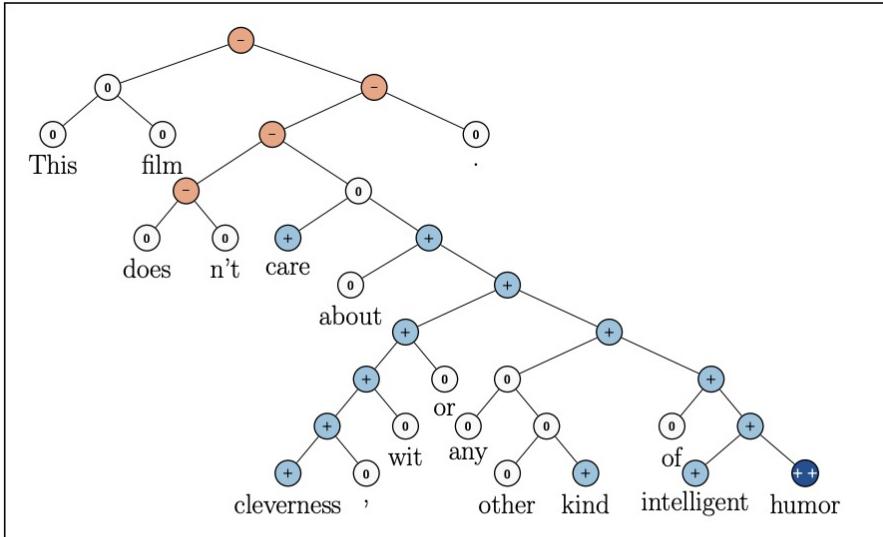


Figure 1: Example of the Recursive Neural Tensor Network accurately predicting 5 sentiment classes, very negative to very positive (--, -, 0, +, ++), at every node of a parse tree and capturing the negation and its scope in this sentence.

Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. EMNLP 2013.
<https://aclanthology.org/D13-1170.pdf>

Manning and Potts recognized by ACL

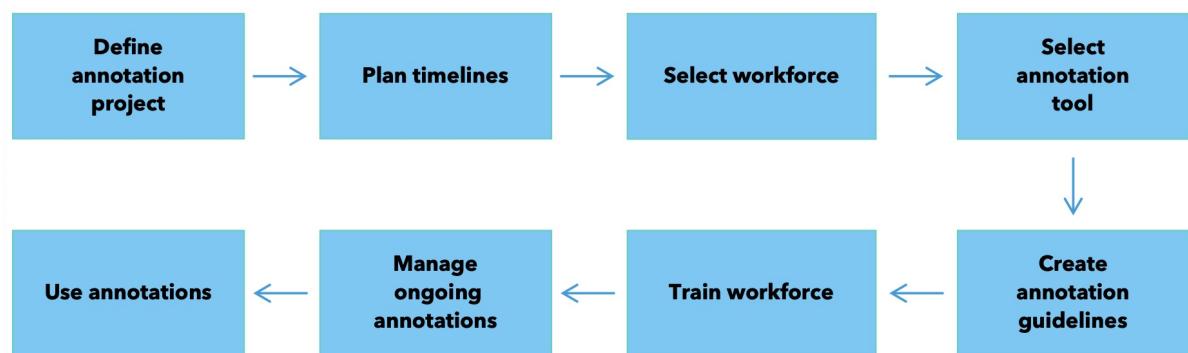
November 10, 2023

[Chris Manning](#) and [Chris Potts](#) are two of the co-authors of a paper that received the [2023 ACL Test-of-Time Paper Award](#), which recognizes four papers for their long-lasting impact on the field of Natural Language Processing and Computational Linguistics. The paper was published in 2013 and is titled “Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank.”

<https://linguistics.stanford.edu/news/manning-and-potts-recognized-acl>

数据标注

Tseng et al. (2020) Best Practices for Managing Data Annotation Projects.



- 需要标注多少
- 采样合适的数据
- 创建标注指南
- 雇佣/监督标注人员
- 评估质量

需要多少测试/开发数据？

- 使不同方法之间有统计学显著差异(如 $p<0.05$)
- 如何估计多少才够？统计效力(Power Analysis) (Card et al. 2020)
 - 对设置之间的效应大小做出假设（例如，测试模型之间的预期精度差异）
 - 给定效应大小，显著性阈值，确定在大多数试验中需要多少数据才能获得显著效果

Card et al. With Little Power Comes Great Responsibility. <https://aclanthology.org/2020.emnlp-main.745.pdf>

需要多少训练数据？

- 多多益善
- 预训练+微调（数据量大大减少？）
- 智能数据选择 - 主动学习？

如何采样数据？

- 覆盖关心的**领域** (Domain)
- 语言，用户人群
- 文档：**NLP的数据声明** (Bender & Friedman 2018)
 - 数据有偏见、系统就有偏见

Curation Rationale	Speech Situation
Language Variety	Text Characteristics
Speaker Demographic	Recording Quality
Annotator Demographic	Other Comments

Emily M. Bender, Batya Friedman: Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. Trans. Assoc. Comput. Linguistics 6: 587-604 (2018)

标注规范

汽车领域事件抽取标注规范

一、 事件类型和事件元素

事件的标注包括事件的类型的事件的元素；

事件类型（8类）：

企业负面、 获取奖项、 代言合作/解约、 新品发布、 事故灾害、 会议活动、 用户投诉、 涉事主体回应；

事件元素有（5类）：

涉事主体、 谓语、 宾语、 时间状语、 地点状语； 其中涉事主体是必须的， 其他是非必须的。

二、 关注的事件

标注的事件是跟世界、 国家、 公司/组织、 品牌、 明星相关的事件， 其中产品营销事件、 股价相关事件、 历史事件、 不明确主宾的事件不标注； 如：

- 1、 双十一**手机大减价； （产品营销事件， 不标注）
- 2、 **股价昨天大降 14.7%； （股价相关事件， 不标注）
- 3、 秦始皇统一六国； （历史事件， 不标注）
- 4、 某明星中止与某公司合作； （不明确主宾， 某明星不知道是谁， 某公司也不知道是哪个公司， 不标注）

招聘标注人员

- **你**: 小型项目
- **同僚**: 朋友或其他学生/同事
- 线上:
 - **自由职业者**: 通过UpWork这样的网站
 - **众包**: 通过像Mechanical Turk这样的网站
- 规模: 从小到大!
- 注意: 对于主观任务, IRB批准可能是必要的

标注质量评估

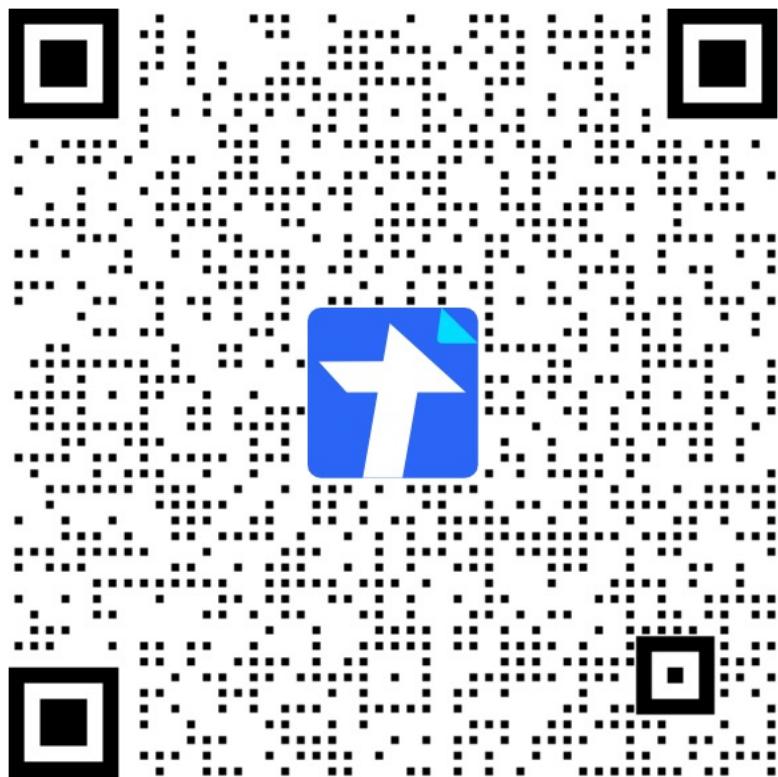
- 标注表现(准确性/BLEU/ROUGE)：双盲标注，测量指标
- κ 统计(Carletta 1996)：

$$\kappa \equiv \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - \boxed{p_o}}{1 - \boxed{p_e}}$$

Observed agreement
Expected agreement

- 如果低，你可能需要：
 - 重新修改标注指南
 - 雇佣更好的标注人员
 - 重新思考任务是否可行

In-class Practice II



- 在自然语言处理的情感标注任务中，2名标注员对 100 条文本进行 “正面(P)/负面(N)” 二分类标注。标注结果如下：
 - 两人均标为 “正面”：35 条；
 - 两人均标为 “负面”：40 条；
 - 标注员 1 标 “正面”、标注员 2 标 “负面”：15 条；
 - 标注员 1 标 “负面”、标注员 2 标 “正面”：10 条。
- 请根据 Carletta (1996) 的 κ 统计公式，计算两名标注员的 κ 系数，并判断标注一致性水平（参考标准：($\kappa > 0.8$)为优秀， $(0.67 < \kappa \leq 0.8)$ 为良好， $(\kappa \leq 0.67)$ 为需改进）

标注案例1：细粒度情感分析

- 任务：将输入文本转换成相应四元组
 $\mathbf{x} \rightarrow \{(c, a, o, s)\}$
- aspect category, aspect term, opinion term, sensitive polarity
- Ex. received the express package on time, work very speedy, but touch screen not sensitive
(Logistics#Speed, express package, NULL, POS)
(Performance#Run Speed, NULL, very speedy, POS)
(Screen#Sensitivity, touch screen, not sensitive, NEG)

2.2 Annotation

A team of professional labelers is asked to label the texts following the guidelines in Appendix A.2. Two annotators individually annotate the same sample by our internal labeled system. The strict quadruple matching F1 score between two annotators is 77.23%, which implies a substantial agreement between two annotators (Kim and Klinger, 2018). In case of disagreement, the project leader

A Unified One-Step Solution for Aspect Sentiment Quad Prediction. <https://aclanthology.org/2023.findings-acl.777.pdf>

标注案例2：情感对话生成

- Input: $\mathcal{D} = \{(\mathbf{x}_i, c_i, \mathbf{y}_i)\}_{i=1}^N$
- Output: $p(\mathbf{y}|\mathbf{x}, c)$

Post The match is over. England 2-1 Denmark.

R_H **Congratulations** England, keep going in the next match!

R_{HI} **Haha**, is England going to win the champion?

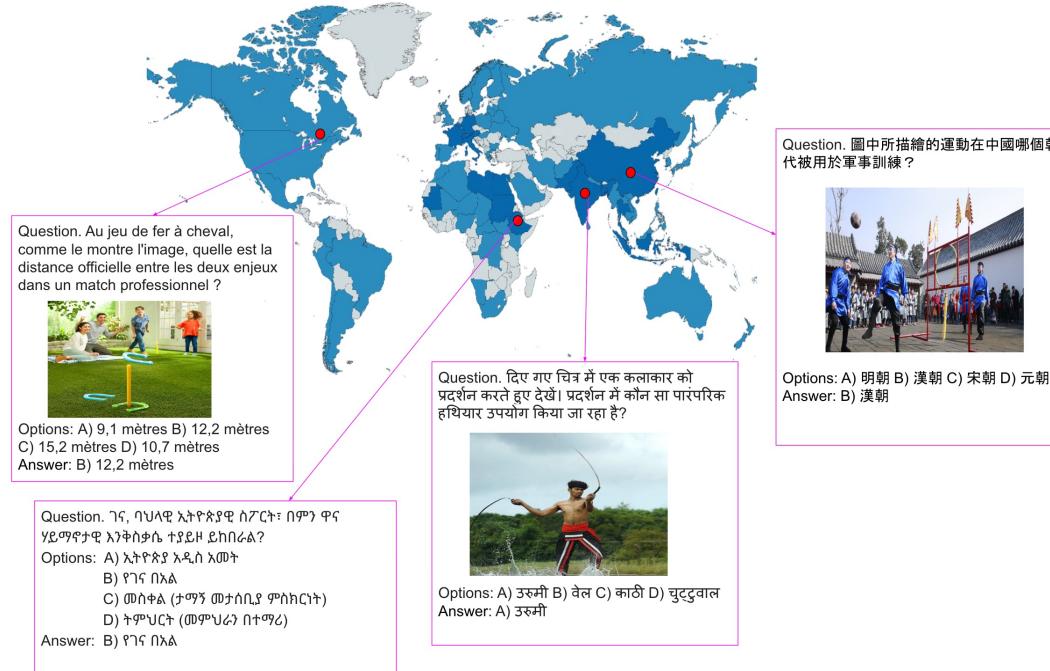
R_{HIL} **Haha**, do not know why this time I am still very happy?

• 实验结果

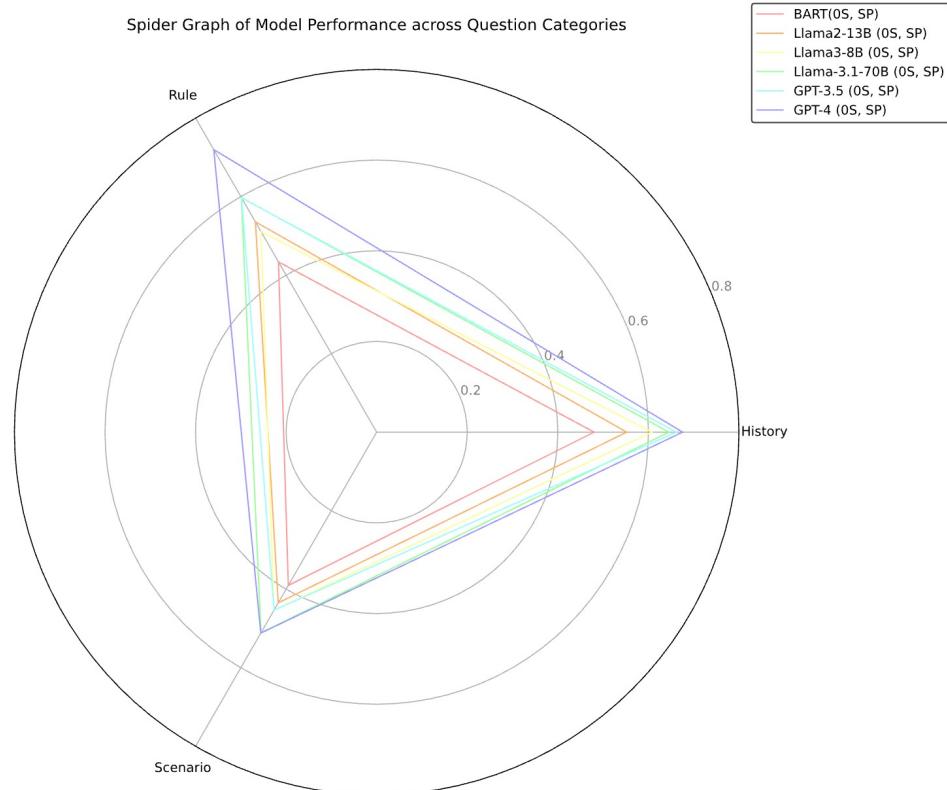
Method	Relevance				Diversity (%)		Human Evaluation			Len. Avg.± Std.
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Dist. 1	Dist. 2	Quality	Good	Accept	
MMI-bidi	10.26	6.03	4.18	3.20	0.15	2.61	1.56±0.57	16%	48%	11.3±1.9
SC-S2S	9.63	5.92	3.87	2.95	0.17	2.47	1.50±0.55	18%	48%	10.1±2.2
DCVAE	11.72	6.97	4.82	3.88	0.18	3.07	1.73±0.94	24%	54%	9.1±2.6
T2T	17.11	9.26	6.16	4.63	0.13	2.81	1.83±0.90	24%	56%	12.0±1.9
PHED _H	16.75	9.03	6.00	4.43	0.14	3.72	2.03±0.90	42%	66%	12.9±3.1
PHED _S	17.69	9.69	6.55	4.95	0.13	4.26	1.98±0.89	39%	58%	14.4±3.8
PHED _{HD}	17.04	9.17	6.08	4.51	0.14	3.74	2.08±0.79	40%	70%	13.3±3.1
PHED _{HI}	15.61	8.11	5.21	3.78	0.12	2.96	2.01±0.82	33%	54%	13.6±3.2
PHED _{SD}	18.28	9.82	6.57	4.94	0.13	4.26	2.02±0.86	39%	56%	14.6±3.8
PHED _{SI}	14.43	8.02	5.44	4.14	0.13	3.64	1.96±0.88	30%	61%	13.6±3.5
PHED _{HDS}	14.26	7.56	4.95	3.64	0.15	3.52	2.01±0.96	41%	65%	11.2±1.5
PHED _{HDL}	18.69	9.82	6.37	4.65	0.11	3.62	1.95±0.95	35%	64%	15.6±3.1
PHED _{HIS}	12.48	6.47	4.12	2.99	0.14	2.92	2.10±0.82	31%	71%	11.3±2.1
PHED _{HIL}	16.67	8.53	5.47	3.88	0.11	2.90	1.93±0.86	25%	59%	16.0±3.0
PHED _{SDS}	14.47	8.07	5.41	4.21	0.16	4.61	1.87±1.06	35%	54%	11.3±1.9
PHED _{SDL}	18.06	9.53	6.28	4.65	0.15	4.06	1.86±0.99	31%	54%	16.8±3.3
PHED _{SIS}	11.86	6.67	4.67	3.59	0.13	3.83	1.81±0.97	28%	53%	11.2±1.8
PHED _{SIL}	15.88	8.57	5.69	4.52	0.11	3.53	1.77±0.89	22%	47%	16.4±3.2

Progressive Open-Domain Response Generation with Multiple Controllable Attributes. IJCAI 2021

标注案例3：Multilingual, Multicultural Benchmark for Sports



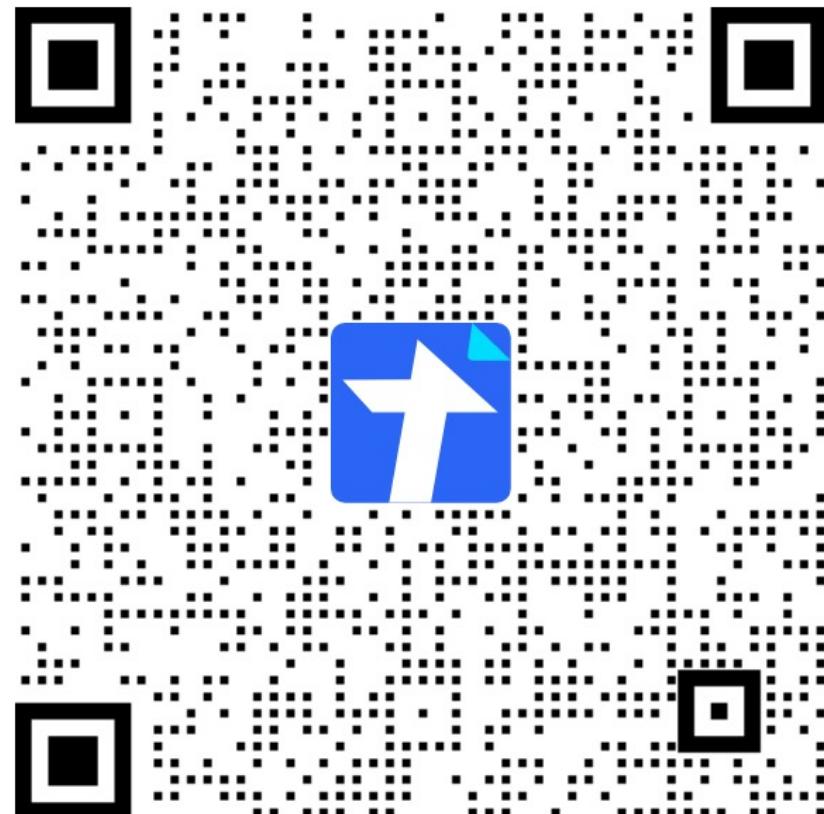
Spider Graph of Model Performance across Question Categories



Let's Play Across Cultures: A Large Multilingual, Multicultural Benchmark for Assessing Language Models' Understanding of Sports. EMNLP'25. <https://arxiv.org/submit/6822047/view>

一句话总结

- 复习NLP系统通用框架和流程
 - 三组数据
- 数据处理
 - C4清洗过程
 - ...
- 大模型评测体系
 - 评测维度
 - 评测方法
 - 评测基准
- 数据标注流程与规范
 - 规范(确定标注量、采样数据、创建指南、评估质量)
 - κ 值评估一致性



Ans. to In-class Practice I

- 已知 $R_A = 1400, R_B = 1500, K = 32$

1. A胜的期望是: $E_A = \frac{1}{1+10^{(R_B-R_A)/400}} = \frac{1}{1+10^{0.25}} \approx \frac{1}{1+1.7783} \approx 0.3599$

2. 根据Elo更新公式: $R'_A = R_A + K \cdot (S_A - E_A)$

- A获胜, $S_A = 1, R'_A \approx 1400 + 32 * (1 - 0.3599) = 1420.5$

- B获胜, $S_A = 0, R'_A \approx 1400 + 32 * (0 - 0.3599) = 1388.5$

- 双方平局, $S_A = 0.5, R'_A \approx 1400 + 32 * (0.5 - 0.3599) = 1404.5$

3. 分析上述三种情况

- A胜则涨20.5分, B胜B则只涨11.5(因A的失分是 $1400 - 1388.5 = 11.5$), 双方平局时A涨4.5分
- 该结果反映, 弱者爆冷得分更多

Ans. to In-class Practice II

- 已知

	标注员2: 正面	标注员2: 负面
标注员1: 正面	35	15
标注员1: 负面	10	40

- 计算

- 实际一致率 p_o : $p_o = \frac{\text{实际一致样本数}}{\text{总样本数}} = \frac{35+40}{100} = 0.75$

- 期望一致率 p_e

- 标注员1: 标“正面” / “负面”比例: $p_1(P) = \frac{35+15}{100} = 0.5, p_1(N) = 0.5$

- 标注员2: 标“正面” / “负面”比例: $p_2(P) = \frac{35+10}{100} = 0.45, p_2(N) = 0.55$

- $p_e = p_1(P) * p_2(P) + p_1(N) * p_2(N) = 0.5$

- κ 值:

- $\kappa = \frac{p_o - p_e}{1 - p_e} = \frac{0.75 - 0.5}{1 - 0.5} = \frac{0.25}{0.5} = 0.5$

- 因为 $\kappa < 0.67$, 因此需要改进