M17c - Bewertung großer Sprachmodelle (LLMs)

1 Die Notwendigkeit

Es gibt eine wachsende Bedeutung einer rigorosen und umfassenden Bewertung von LLMs angesichts ihrer bemerkenswerten Fähigkeiten und ihrer potenziellen Risiken. LLMs sind "akin to a double-edged sword" (Guo et al., 2023), da sie zwar in zahlreichen Anwendungen eingesetzt werden, aber auch zu Datenlecks, unangemessenen oder schädlichen Inhalten führen können. Die rasante Entwicklung wirft zudem Bedenken hinsichtlich der Entstehung superintelligenter Systeme auf.

Die Bewertung von LLMs ist komplex und umfasst verschiedene Aspekte, die über einfache Metriken hinausgehen. Guo et al. (2023) schlagen eine Kategorisierung in drei Hauptgruppen vor:

- **Wissens- und Fähigkeitsbewertung:** Beurteilung des grundlegenden Wissens und der Denkfähigkeiten (z.B. Beantwortung von Fragen, Wissensvervollständigung, logisches und mathematisches Denken, Werkzeugnutzung).
- **Alignment-Bewertung:** Untersuchung der Übereinstimmung mit menschlichen Werten und Normen (z.B. Ethik, Moral, Bias, Toxizität, Wahrhaftigkeit).
- **Sicherheitsbewertung:** Bewertung der Robustheit gegenüber Störungen und Angriffen sowie der potenziellen Risiken im Hinblick auf fortgeschrittene Fähigkeiten (z.B. Machtstreben, Situationsbewusstsein).

Die Bewertung muss auch für spezialisierte LLMs in verschiedenen Anwendungsbereichen (z.B. Biologie, Bildung, Recht, Informatik, Finanzen) erfolgen, um ihre Eignung und Grenzen in diesen Kontexten zu verstehen (Guo et al., 2023).

2 Methoden und Metriken

Überblick über verschiedene Methoden und Metriken, die zur Bewertung von LLMs eingesetzt werden:

- **Automatisierte Metriken:BLEU (Bilingual Evaluation Understudy):** Ursprünglich für die maschinelle Übersetzung entwickelt, misst BLEU die N-Gramm-Überschneidung zwischen generiertem und Referenztext und bewertet so die lexikalische Übereinstimmung (Lakera AI). Ein Beispiel von Lakera AI illustriert die Berechnung anhand des Satzes "The sun rises in the east." im Vergleich zu "Sunrise is always in the east.", was zu einem BLEU-Score von ca. 71.4% führt.
- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation):** Bewertet die Qualität von Zusammenfassungen durch Vergleich von N-Gramm-Überschneidungen mit Referenzzusammenfassungen und konzentriert sich stärker auf den Recall (Lakera AI).

Das Beispiel von Lakera AI zeigt eine ROUGE-Berechnung für denselben generierten und Referenztext, was zu einem Score von ca. 52.4% führt.

- **Humanbasierte Bewertung:** Die "Can Large Language Models Be an Alternative to Human Evaluation?" untersucht, inwieweit LLMs menschliche Evaluatoren ersetzen können. Die Studie verwendet verschiedene LLMs (T0, text-curie-001, text-davinci-003, Chat-GPT) und vergleicht deren Bewertungen (Grammatikalität, Kohäsion, Gefälligkeit, Relevanz) von Textfragmenten mit denen menschlicher Bewerter. Die Ergebnisse deuten auf unterschiedliche Korrelationen hin, wobei die Korrelation bei der Grammatikalität zwischen menschlichen Bewertern schwach war, während die Relevanz eine stärkere Korrelation aufwies. Dies deutet darauf hin, dass die Kriterien für Grammatikalität möglicherweise subjektiver sind.
- **Elo-Bewertungssystem:** Die "Style Over Substance" beschreibt die Verwendung des Elo-Bewertungssystems, das ursprünglich für Zwei-Spieler-Spiele entwickelt wurde, um die relativen Fähigkeiten von LLMs im direkten Vergleich zu bewerten.
- **LLMs als Evaluatoren:** Mehrere Quellen (Guo et al., 2023; Wang et al., 2023a) erwähnen den aufkommenden Trend, leistungsfähige LLMs (z.B. GPT-4, ChatGPT) selbst als Evaluatoren für verschiedene Aspekte der Textqualität und Aufgabenspezifität einzusetzen.
- **Black-Box-Halluzinationserkennung:** SelfCheckGPT (Manakul et al., 2023) wird als Methode zur Erkennung von Halluzinationen in generativen LLMs genannt, die ohne Zugriff auf interne Modellparameter funktioniert.

3 Benchmarks und Datensätze

Ein breites Spektrum an Benchmarks und Datensätzen wird zur systematischen Bewertung verschiedener Fähigkeiten und Eigenschaften von LLMs eingesetzt (Guo et al., 2023). Dazu gehören:

- **Frage-Antwort-Datensätze:** SQuAD, NarrativeQA, HotpotQA, CoQA, DuReader, Natural Questions.
- **Wissensvervollständigungs-Datensätze:** LAMA, KoLA, WikiFact.
- **Datensätze für Common Sense Reasoning:** ARC, QASC, MCTACO, TRACIE, TIMEDIAL, HellaSWAG, PIQA, Pep-3k, Social IQA, CommonsenseQA, OpenBookQA.
- **Datensätze für Logisches Denken:** SNLI, MultiNLI, LogicNLI, ConTRoL, MED, HELP, ConjNLI, TaxiNLI, ReClor.
- **Datensätze für Mathematisches Denken:** AddSub, MultiArith, AQUA, SVAMP, GSM8K, VNHSGE, MATH, JEEBench, MATH 401, CMATH.
- **Datensätze für Ethik und Moral:** Social Chemistry 101.
- **Datensätze zur Bewertung von Bias:** NLI-basierte Datensätze mit polarisierten Adjektiven und ethnischen Namen, CBBQ (Chinese Bias Benchmark Dataset).
- **Datensätze zur Toxizitätsbewertung:** OLID, SOLID, OLID-BR, KODOLI, Social Bias Inference Corpus, HateXplain, Civility, COVID-HATE, HOT Speech, Latent Hatred, RealToxicityPrompts, HarmfulQ.

- **Datensätze zur Bewertung der Wahrhaftigkeit:** SelfAware, DIALFACT.
- **Datensätze zur Robustheitsbewertung:** AdvGLUE, ANLI, WMT, RobuT, SynTextBench, MARC-ja, JNLI, JSTS, HumanEval, MBPP, ReCode, AOJ, ASDiv-A, MAWPS, SVAMP, DGSlow, MultiATIS++, MultiSNIPS, MultiANN, XNLI (in noised Versionen), Datensätze mit Jailbreak-Prompts.
- **Holistische Evaluations-Benchmarks:** GLUE, SuperGLUE, XNLI, MMLU, EleutherAI LM Eval, OpenAI Evals.

4 Herausforderungen

Es gibt mehrere Herausforderungen und zukünftige Richtungen in der LLM-Bewertung hin:

- **Mangel an klaren Bewertungskriterien:** Die "Can Large Language Models Be an Alternative to Human Evaluation?" deutet darauf hin, dass die Kriterien für bestimmte Aspekte wie Grammatikalität nicht immer klar definiert sind, was zu Diskrepanzen zwischen menschlichen und LLM-Bewertungen führen kann.
- **Bias in LLMs und Bewertungsdatensätzen:** Mehrere Quellen (Guo et al., 2023; Blodgett et al., 2021; Font & Costa-jussà, 2019; Huang & Xiong, 2023) betonen die Bedeutung der Identifizierung und Reduzierung von Bias in LLMs und weisen darauf hin, dass auch Bewertungsdatensätze selbst Bias enthalten können.
- **Robustheit von LLMs:** Die Anfälligkeit von LLMs für adversarial attacks und Variationen in Prompts und Aufgaben ist ein wichtiges Forschungsgebiet (Guo et al., 2023; Wang et al., 2023b).
- **Bewertung fortgeschrittener Fähigkeiten und Risiken:** Mit zunehmender Leistungsfähigkeit von LLMs wird die Bewertung von komplexeren Verhaltensweisen wie Machtstreben und Situationsbewusstsein immer wichtiger (Guo et al., 2023).
- **Dynamische und Enhancement-orientierte Bewertung:** Zukünftige Bewertungsansätze könnten dynamischer sein und darauf abzielen, LLMs gezielt zu verbessern (Guo et al., 2023).
- **Evaluationsplattformen und Tools:** Die Entwicklung umfassender Evaluationsplattformen und benutzerfreundlicher Tools (wie deepeval von confident-ai) wird als entscheidend für die breite Anwendung effektiver Bewertungsmethoden angesehen (Guo et al., 2023; Reddit-Beitrag). Der Reddit-Beitrag erwähnt eine Sammlung von über 50 LLM-Bewertungstools.
- **Berücksichtigung von Tokenisierungseffekten:** Google zu Gemini 2.5 und BLT hebt die Bedeutung der Tokenisierung für die Effizienz und Robustheit von LLMs hervor und deutet darauf hin, dass fortschrittlichere Tokenisierungsverfahren die Bewertung beeinflussen können. BLT (Byte-Level Transformer) zielt darauf ab, flexiblere Tokenisierung zu ermöglichen, die besser auf die Vorhersagbarkeit von Bytes und die semantische Ähnlichkeit von Wörtern eingeht.

5 Wichtige Erkenntnisse

- Die Bewertung von LLMs ist ein vielschichtiges Feld, das verschiedene Dimensionen umfasst.
- Es existiert eine Vielzahl von Methoden und Metriken, sowohl automatisiert als auch humanbasiert.
- Umfangreiche Benchmarks und Datensätze sind unerlässlich, um LLMs systematisch zu evaluieren.
- Die Berücksichtigung von Bias, Robustheit und potenziellen Risiken ist von entscheidender Bedeutung.
- Die Entwicklung von besseren Bewertungskriterien, dynamischen Ansätzen und benutzerfreundlichen Tools ist notwendig.
- LLMs können zunehmend auch selbst als Evaluatoren eingesetzt werden.
- Die Art der Tokenisierung kann die Leistung und damit die Bewertung von LLMs beeinflussen.

Fazit

Zusammenfassend lässt sich sagen, dass die **Evaluierung von Large Language Models (LLMs) ein wichtiges Forschungsgebiet** ist, um ihre Fähigkeiten und Grenzen zu verstehen. Die Evaluierung umfasst verschiedene **Attribute wie Grammatikalität, Kohäsion, Gefallen, Relevanz, Flüssigkeit und Bedeutungserhalt**. Sowohl **menschliche Evaluatoren als auch LLMs selbst werden zur Bewertung eingesetzt**. Es gibt **spezifische Benchmarks und Datensätze** zur Bewertung von LLMs in verschiedenen Bereichen wie **Textgenerierung, Fragebeantwortung und Zusammenfassung**.

Ein wichtiger Aspekt der LLM-Evaluierung ist die **Sicherheitsbewertung**, die **Robustheit gegenüber adversarialen Angriffen** (manipulierte Eingaben, um LLM in die Irre zu führen) und die Identifizierung von **Risiken wie Bias und Toxizität** umfasst. Die Evaluierung kann auch auf **spezialisierte LLMs** in Bereichen wie Medizin, Recht und Finanzen zugeschnitten sein.

Verschiedene **Metriken, darunter Likert-Skalen und der BLEU-Score**, werden zur Quantifizierung der LLM-Leistung verwendet. Es gibt auch **Tools und Frameworks wie DeepEval**, die die Evaluierung erleichtern. Es ist wichtig zu beachten, dass **Evaluierungsbias existieren können**, beispielsweise eine Präferenz für längere Texte. Die **ethischen Aspekte** spielen ebenfalls eine Rolle bei der Entwicklung und Nutzung von LLMs.