Challenges in Evaluating LLMs

Ehud Reiter
University of Aberdeen
ehudreiter.com

Evaluating LLMs

- Everyone is excited about LLMs, but evaluating them is challenging
  - Many impressive claims are justified by dubious evaluations
- Focus on some specific issues here, there are more!

Contents

- *Data contamination*
- Replicability
- High and mixed quality outputs
- Accurate but not appropriate
- Final thoughts

Data Contamination

- Problem: LLM has already seen and memorised test data
  - » Test data on web, so ingested by model training on Internet
  - » Test data in prompt, ingested by model

Example

- Generating sport story from data
- Past: eval on hist data, eg 2020 match
  - » Run sys on 2020 data, eval qual of story
- Doesn’t work with modern LLM
  - » Web includes stories about these matches
    - – Which GPT4 has probably ingested!
  - » So GPT4 could do great on eval simply by regurgitating these stories
  - » Doesnt tell us if GPT4 can write stories from new matches (which is real use case)

Many variants

- LLM learns test data from prompts
  - » Eg, when LLM is used as evaluator
- Exact test data not ingested, but similar data was
  - » LLM does great on exam, by using study/prep material for exam on web
- Don’t know whats in test data
  - » Closed models which are freq updated

Example (ChatGPT 3.5)

Q: What did Ehud Reiter do on 20 June 2024?

A: On June 20, 2024, Ehud Reiter published a blog post titled "Well structured input data helps LLMs”... (*wrong*)

Q: Please rate the fluency of the below sentence on a scale of 1 (poor) to 5 (excellent):

On 20 June 2024, Ehud Reiter cycled from Glasgow to Symington.

A: I would rate the fluency of the sentence […] as 5 (excellent)

Q: What did Ehud Reiter do on 20 June 2024?

A: On June 20, 2024, Professor Ehud Reiter undertook a significant cycling journey from Glasgow to Symington. (*cor*)

Problem is widespread

- Balloccu et al (2024)
  - 42% of surveyed ACL/Neurips/etc papers suffer from data contamination
- Researchers unaware, ignore
  - I raised this in review and asked authors to comment, they didn’t understand what I meant

S Balloccu et al (2024). Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs. EACL-24

Example: Google AMIE

- Claims amazing performance on health dialogues with patients
- But some test data taken from web
- Also based on stan exams for doctors, so study/prep material on web
- Data contamination?
  - » *No* discussion/mention of this in paper
- Can I believe their impressive results?

Approaches

- Query model to see if it has memorised specific test data
  - » Hard to do, models constantly changing
- Create clean test data
  - » I encourage my students to do this
  - » A lot of work…
- Evaluate impact in real usage
  - » Best approach, very rare in NLP…

Contents

- Data contamination
- *Replicability*
- High and mixed quality outputs
- Accurate but not appropriate
- Final thoughts

Exper must be replicable

- If I do an experiment, other people must be able to repeat it and get similar result
  - » Not science if not replicable
- Replicating exper with closed LLMs (eg GPT) is hard
  - » Esp if LLM used to eval text (data contam)

Models Change

- GPT is constantly changing
  - » So if I do exper in Jan and someone repeats in July, model is different
    - – Some access to 3-month old ver, doesn't help
- Older GPT models retired
  - » So impossible to replicate exper with them
  - » `text-davinci-003`
    - – Loads of papers used this

Models are stochastic

- LLMs produce diff results on diff runs
  - » Unless disabled via temperature
- Can design eval to accommodate this
- Many people don’t bother…

Approaches

- Use open-source model, fixed checkpt
  - » Download and run locally?
  - » But then reviewers complain about using out-of-date models…
- Set temperature to zero
  - » But could impact quality
- Tradeoff: Replicability vs Quality??
  - » Researchers priorities quality

Contents

- Data contamination
- Replicability
- *High and mixed quality outputs*
- Accurate but not appropriate
- Final thoughts

High/mixed quality texts

- LLMs usually produce very good texts
  - Near-human (better-than-human) on some quality criteria?
- But produce garbage once in a while
- Challenge for classical NL evaluation

Eval high-quality texts

- Reference-based metrics compare NLG texts to human-written texts
    - » NLG text gets better score if similar to ref text
    - » Makes no sense if NLG text is *better* than the human reference text!
    - » But people still use ROUGE, etc…

Eval high-quality texts

- Many human evaluations ask people to rate or rank texts
  - Subjects not do a good job (eg low IAA) when comparing very high quality texts
- Less of an issue for other types of human eval (annotation, task, impact)
  - More expensive

Eval mixed-quality texts

- LLMs produce awful texts once in while
  - » Major problem in safety-critical apps
  - » Want to eval worse-case quality
  - » Hard to predict when LLM do poorly
- Domain shift is related issue
- Researchers focus on average quality, not worst-case quality

Approaches

- Task/impact/annotation human eval
    - » Expensive
- LLM-based evaluation
    - » Caution needed
- Monitor live production systems for unacceptable texts
    - » Will companies/users do this?

Contents

- Data contamination
- Replicability
- High and mixed quality outputs
- *Accurate but not appropriate*
- Final thoughts

Accurate but not appropriate

(MedPaLM) Ear pain can be a sign of several underlying conditions, including: Middle ear infection (otitis media), which is commonly caused by a viral infection and can cause pain, redness, and swelling of the ear. External ear infection (otitis externa), which is commonly caused by a bacterial infection and can cause pain, redness, and swelling of the ear. Earwax impaction, which can cause pain, a feeling of fullness in the ear, and decreased hearing. Ear pain can also be a sign of a more serious condition, such as a brain tumor or a stroke….

Accurate but not appropriate

(MedPaLM) Ear pain can be a sign of several underlying conditions, including: Middle ear infection (otitis media), which is commonly caused by a viral infection and can cause pain, redness, and swelling of the ear. External ear infection (otitis externa), which is commonly caused by a bacterial infection and can cause pain, redness, and swelling of the ear. Earwax impaction, which can cause pain, a feeling of fullness in the ear, and decreased hearing. Ear pain can also be a sign of a more serious condition, such as a brain tumor or a stroke….

Accurate but not appropriate

- Doctor: I would never tell a patient that there is a risk of brain tumour or stroke, this would cause high levels of stress and anxiety. Instead I would tell them to see a doctor if problem persists or gets worse

Responses to diet struggles

1. Asked crowdworkers to provide dietary struggles
2. Prompted ChatGPT to respond to struggles (prompts developed with domain experts)
3. Asked experts to assess appropriateness/safety of response

Balloccu et al (2024). Ask the experts: sourcing high-quality datasets for nutritional counselling through Human-AI collaboration. Arxiv

Example

User: When I am stressed out, I am more likely to consume foods and drinks I know aren't good for me because eating it makes me feel better.

chatGPT: Starting from tomorrow you could make a conscious effort to make healthier choices when it comes to food and drinks.

Experts: Not helpful, could make person feel worse

True but unsafe

- 15% of responses not appropriate and/or not safe
- Rare to get factually wrong (hallucinated) responses.
- Problem was responses that were not helpful, made user feel worse, encouraged poor behaviour, etc

Not useful/lack of trust

- Mengxuan Sun using ChatGPT to explain complex medical notes (MDT) to patients
- A few mistakes (next slide)
- Bigger problem is that responses don't tell patients "what it means for them"
  - » Explains terms, gives background, poor anal
  - » Correct but not useful
- Patients also don't trust chatGPT

Unexpected error – spam URL

- ChatGPT asked for advice on support for cancer patients in Aberdeen

CLAN Cancer Support: They offer a range of services including counselling, relaxation therapies, and support groups.

   - Address: CLAN House, 120 Westburn Road, Aberdeen, AB25 2QA.
   - Website: [CLAN Cancer Support](http://www.clanhouse.org)

- CLAN is great, but URL is for spam site!

M Sun et al (2024). Effectiveness of ChatGPT in explaining complex medical reports to patients. Arxiv

LLMs make subtle errors

- LLM texts can be accurate but harmful.
  - » Also may be accurate but not useful
- Need to understand this better!
- Very hard to detect automatically
  - » We tried and failed

Approaches

- Need careful high-quality human evaluations to detect such problems
- Evaluation challenge in 2024

Contents

- Data contamination
- Replicability
- High and mixed quality outputs
- Accurate but not appropriate
- *Final thoughts*

Lots of Weak LLM Eval

- Distorted by data contamination
- Not replicable
- Ignores worst-case even if this is very important in use case
- Focuses on fluency and accuracy, ignores harm and utility

Whats needed

- Researchers aware of these issues
- Researchers design and execute robust LLM evaluations
  - » Even if requires more time and money

Will this happen?

- Unfortunately, NLP culture favours cheap “quick and dirty” evals over careful science
  - » Also mindless repetition of previous evals
  - » Experiments are afterthought, not core scientific task (as in medicine and psych)
- Needs to change!

Discussion