

Evaluation: Concepts

Ehud Reiter

Evaluation

- Assess how well a system/model works
 - Where it needs to be improved
 - How it compares to other systems
- Usually measure key quality criteria for texts produced system

Experimental evaluation

- We use *experiments* to measure quality criteria
- Experiments must be carefully designed, executed, and analysed
- Poor experiments tell us nothing

Discussion

- What evaluations have people done?
- What worked well?
- What went wrong?

Contents

- *Types of evaluation*
- Hypothesis testing
- Experimental design
- Replication
- Challenges in evaluating LLMs

Types of evaluation

- Human evaluation – ask people to assess quality criteria
 - Or assess impact on task performance
- Automatic evaluation
 - Use metrics, algorithms, models to assess quality criteria
- Impact evaluation
 - Measure real-world impact on KPIs (key performance indicators)
- Discussed in detail in later talks
- Quick summary here

Human evaluation

- *Rating/Ranking*: Ask subjects to read texts and estimate quality
 - Other use Likert scale (“This text is accurate”)
 - Can also ask subjects to rank several texts based on quality criteria
 - Each rating/ranking for one quality criteria
- *Annotation*: Ask subjects to mark up specific problems
 - Incorrect statements, difficult to understand phrases, etc
 - More objective, higher subject agreement, better guidance on what to fix
- *Task-based*: Ask subjects to do a task with the text
 - Most common is check and fix a draft text; how long does this take?
 - More meaningful, but depends on UI, subject, context

Automatic evaluation

- *Reference-based*: Measure similarity of generated text to high-quality “reference” text
 - Various algorithms, best use trained or fine-tuned models
- *Referenceless*: Directly measure a quality criteria
 - Best approach is to prompt an LLM (eg, GPT4) to do this

Impact evaluation

- Deploy a system in real-world, and assess impact on important KPI (key performance indicators)
 - Medical diagnosis: does accuracy increase?
 - Report writing: is this faster, fewer mistakes?
 - Behaviour change: how many people are convinced to stop smoking, drive more safely, etc
- Experiment must be ethical!

Contents

- Types of evaluation
- *Hypothesis testing*
- Experimental design
- Replication
- Challenges in evaluating LLMs

Hypothesis testing

- Scientific experiments test *hypotheses*
 - Smokers are more likely to get lung cancer
 - Nothing can travel faster than the speed of light
 - The moon is made of green cheese
- Different techniques are used
 - Measure lung cancer rates in smokers and non-smokers
 - Try to accelerate particles faster than light
 - Get astronauts to bring back samples of the moon

Hypothesis testing

- Can disprove (falsify) hypotheses
 - Moon is not made of green cheese
- Very hard to completely prove hypotheses, because of confounds
 - Maybe speed of light exceeded in weird context we have not checked
 - Maybe there is an unknown factor which both encourages smoking and causes cancer

NLG hypotheses

- Texts from System A are better than texts from System B under quality criteria C
 - Ehud's sportswriter produces more accurate stories than Jane's sportswriter
- Texts from System A achieve a value of at least V on criteria C
 - Average report will take less than one minute for a person to check and edit

Statistical hypothesis testing

- Use statistical tests to check whether a result is genuine or could just be noise.
- Typically compute the probability that the observed results could be seen if the hypothesis is false (*p-value*)
- Experiment only supports results if $p\text{-value} < 0.05$
 - Conventional threshold

Statistical tests

- *Compare means* (eg, A texts better than B texts on criteria C)
 - T-test (comparing two systems)
 - ANOVA (comparing more than two systems)
 - Many others, eg non-parametric
- *Compare categories* (eg, doctors seeing A texts are more likely to make the right diagnosis than doctors seeing B texts)
 - Chi-square
- *Correlations* (eg, longer texts have more mistakes)
 - Pearson correlation
 - Spearman correlation (my preference)
- Many others

Statistical testing

- Huge literature on this, will not discuss more here
- Most universities have classes on this
 - Try medicine or psychology if not in CS
- Important, do learn this!

Contents

- Types of evaluation
- Hypothesis testing
- *Experimental design*
- Replication
- Challenges in evaluating LLMs

Experimental design

- What hypothesis are we testing?
- What evaluation technique will we use?
- What material will be evaluated (test set_
- What humans subects will we use (for human eval)?
- What is the detailed experimental procedure?
- How will data be analysed?

Example: Sportswriter

- *Hypothesis*: Texts from Ehud's sportswriter have fewer errors than texts from Jane's sportswriter
- *Technique*: Human subjects annotate (mark up) errors
- *Material*: 100 texts from each system, produced from data from 100 randomly-selected basketball games
- *Subjects*: Students, who are paid minimum wage
- *Procedure*: Annotation process (see later lecture)
- *Analysis*: Paired t-test comparing per-text error counts

Experimental design and execution

- Evaluation experiment must be carefully designed
 - Sensible hypothesis
 - Appropriate evaluation type and technique
 - Representative test set which model has not seen
 - Subjects who take task seriously (for human evaluation)
 - Experiment is replicable
- Experiments must be carefully executed
 - Avoid code bugs, procedural errors
 - Correctly analyse and report data
 - Respond to questions from readers

Hypothesis mistakes: Measure wrong thing

- The experiment must measure important quality criteria!
- Hypothesis is what is important, not what is easy to measure
- Example: Measuring readability instead of accuracy in a medical application
 - Accuracy is much harder to measure (may need help from doctors)
 - But it is much more important than readability

Hypothesis mistake: Weak baseline

- Compare against state-of-art, not 5-year old system
- Better than old system is not good hypothesis
- Example: GPT3 paper compared against a 5-year old MT system
 - Not meaningful!

Eval technique mistake: outdated evaluation

- Do not use obsolete evaluation techniques such as BLEU or ROUGE
- Much better metrics exists
- Even in 2024, I frequently see papers that use BLEU and ROUGE

Material mistake: Unrepresentative test set

- The test set used in the evaluation must be real data which is representative of actual use
- I know getting good data is a pain, but it is important
- Example: Many medical AI/NLG systems use MIMIC, which is a dataset from ICU in one high-end US hospital.
 - NOT representative of hospitals in general (certainly not UK NHS)
 - NOT representative of other hospital units
 - So don't claim results on MIMIC are general!
- *Many* other examples of this

Material mistake: Data contamination

- The model being evaluated should not have been trained on the test data!
- Fundamental principle of machine learning
- Often violated with LLMs (discussed in later)
- One reason why LLMs evaluations don't translate into real-world effectiveness

Subject mistake: Subjects not representative

- Don't ask CS students to evaluate
 - clinical NLG system (get doctors to do this)
 - system aimed at general public (get broad range of subjects)
- Subjects should be representative of real users

Procedure mistake: Detect lazy subjects

- Many human evaluations use crowdworkers
 - They may click randomly (or using ChatGPT to respond)
 - After all this maximises their income
 - Need to check for this!
-
- Ive seem many experiments with dubious data because of this

Analysis mistake: No stats

- Compute statistical significance of results!!
- Lots of NLP papers do not do this

Execution mistake: Code/reporting errors

- A lot of experiments use buggy code to run the experiment
 - Also bugs in analysis code
 - Distorts results
- Many papers show wrong results
 - Numbers in paper don't match experimental data
 - Sloppiness, not fraud
- Be careful when you execute an experiment!
- Discussed later

Replication mistake: Experiment not replicable

- Scientific experiments must be replicable by other researchers
 - Otherwise not science
- Discussed later
- Example: Experiment using GPT3 text-davinci3 model
 - Model decommissioned, no longer exists
 - So impossible to replicate experiments
 - But I still see papers in 2024 based on this model

Discussion

- Have you seen any of these mistakes
 - Measuring unimportant things
 - Weak baseline
 - Outdated evaluation technique
 - Unrepresentative test set
 - Data contamination
 - Unrepresentative subjects
 - Human subjects don't care
 - Code/analysis bugs
 - Experiment not replicable

Contents

- Types of evaluation
- Hypothesis testing
- Experimental design
- *Replication*
- Challenges in evaluating LLMs

Replication

- Science must be reproducible!
- If I do an experiment, you should be able to repeat the experiment and get similar results
 - Otherwise not science...

ReproHum project

- Explore reproducibility of NLP human eval
 - <https://reprohum.github.io/>
- 20 partner labs reproduce a selected set of NLP evaluations
 - Eval published in ACL or TACL papers
 - Each eval reproduced twice

Depressing Findings

- Most authors are reluctant to help in reproducing their papers
- All papers we have reproduced have flawed execution
- Some experiments could not be replicated

A Belz, et al (2023). Missing Information, Unresponsive Authors, Experimental Flaws: The Impossibility of Assessing the Reproducibility of Previous Human Evaluations in NLP. *Proc of The Fourth Workshop on Insights from Negative Results in NL*

C Thomson, E Reiter, A Belz (2024). Common Flaws in Running Human Evaluation Experiments in NLP. *Computational Linguistics*.

Problem 1: Authors

- In most cases, much easier to reproduce experiments if authors help
 - Provide details of data, exper design, UI, subject recruitment, statistical packages, etc
- Problem: Few authors will provide experimental details

Author Response

- When asked for experimental details
 - 61% of authors did not response
 - 26% responded to initial email but did not provide necessary details
 - **Only 13%** of authors responded and provided experimental details

Problem 2: Poor execution

- Results reported in paper do not match actual experimental data
- Code bugs distort experiments and analysis
- Outliers dropped ad-hoc
- Poor/confusing experimental UI
- Flawed anonymisation (ethics)

Code bug

[below is anonymised]

```
python prepareMTurkFiles.py
    -gold data/gold.txt \
    -baseline data/baseline.txt \
    -sys1 data/sys1.txt \
    -sys2 data/sys2.txt \
    -sys3 data/sys2.txt \
    -output_file outputs/mturk.csv
```

Code bug

[below is anonymised]

```
python prepareMTurkFiles.py
    -gold data/gold.txt \
    -baseline data/baseline.txt \
    -sys1 data/sys1.txt \
    -sys2 data/sys2.txt \
    -sys3 data/sys2.txt \
    -output_file outputs/mturk.csv
```


How common are code bugs?

- After high-qual commercial quality assurance, code has 1-2 errors/100 lines
- Most research/PhD code does not go through commercial quality assurance...
 - So more than 1-2 bugs/100 lines
- I suspect *big* problem in research code
 - Which no one talks about...
- We've seen other examples

Advice: Sanity-Check Results

- Check actual evaluation outputs (not just statistics and scores)
 - Look for bizarre patterns
 - Sys2 and sys3 have identical results
- Investigate if found
 - Either bug (embarrassing if published)
 - Or unexpected discovery!

Execution errors

- In ReproHum Phase 1, our partners reproduced experiments from 5 papers
 - All from ACL or TACL
- ***Every*** paper had execution flaws
- Not a good sign...

Problem 3: results could not be replicated

- One replicated experiment evaluated accuracy in a sportswriting domain by asking Turkers to count incorrect facts
 - Protocol proposed by Wiseman et al (2017)
 - Used by dozens (100s??) papers subsequently
- NOT replicable

Replication of old protocol

Paper	Note	mean errors
Pudupully and Lapata 2021	Original paper	0.07
Gonzalez-Corbelle et al (2023)	Replication 1	0.66
Watson and Gkatzia (2023)	Replication 2	1.525
Watson and Gkatzia (2023)	Replication 2 with academic evaluators	0.0625
Thomson et al (2023)	Better methodology	0.5

Mean number of errors in 4-sentence extract from corpus text

Protocol doesn't work

- Replications give very different results
 - Experiment cannot be trusted!
- Poor design
 - Task is complex and not well defined
 - Subjects not motivated, lack expertise
- Not a good evaluation protocol

<https://ehudreiter.com/2023/10/31/a-bad-way-to-measure-hallucination/>

What matters for replicability?

- Asked ReproHum partners what factors most influence reproduce and reliability
- Most common factor mentioned was guidelines and training for subjects
 - Subjects in human evaluations need to understand what they are supposed to be doing!

Lessons from ReproHum

- Most researchers don't cooperate with replication attempts (*poor* culture)
- Many flaws in executing human eval, such as code bugs
- Some experimental results cannot be replicated
- Subjects need to understand the task (guidelines and training)
- Do better in your evaluations!

Contents

- Types of evaluation
- Hypothesis testing
- Experimental design
- Replication
- *Challenges in evaluating LLMs*

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!

Some challenges

- *Data contamination*
- Replicability
- High and mixed quality outputs
- Subtle errors

Challenge 1: 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

- Sportswriter: Generates sport story from data
- Past: eval on hist data
 - Choose 2018 match not in training data
 - Run sys on 2018 match, eval qual of story
- Doesn't work with modern LLM
 - Human-written stories about 2018 match are on web
 - Which GPT4 has probably ingested as training data
 - So GPT4 could do well simply by repeating these stories
 - Doesn't tell us if GPT4 can write stories from new data
 - Which is what we actually want to do

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. (*right*)

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 standard exams for doctors, so study/prep material on web
- High risk of 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...

Challenge 2: Replicability

- Replicating experiments with closed LLMs (eg GPT) is hard
- 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 to replicating LLM evaluations

- Use open-source model, fixed checkpoint
 - 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 prioritize quality

Challenge 3: 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 when comparing very high quality texts
- Less of an issue for other types of human eval (annotation, task)

Eval mixed-quality texts

- LLMs produce awful texts once in while
 - Problem if need to guarantee quality criteria always meets minimum
 - Very common in safety-critical apps
 - Hard to predict when LLM do poorly
 - Hard to evaluate worse-case quality
- Domain shift is related issue
- Researchers focus on average quality, not worst-case quality
 - Even if worst-case quality is very important in user requirements!

Approaches

- Task/annotation human eval
- LLM-based evaluation??
 - Caution needed
- Monitor live production systems for unacceptable text

Challenge 4: LLMs make subtle errors

- LLM texts can be accurate but harmful.
 - Also may be accurate but not useful
- Need to understand this better!
 - Past: focused on readability, accuracy, content quality criteria
 - Define quality criteria for more subtle aspects

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 (eg, spam URL)
- 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

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

Need to define and measure new quality crit

- Makes user stressed (or other poor emotional reaction)
- Encourages poor behaviour
- Trust

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

Discussion: LLM Evaluations

- Have people seen LLM evaluations which they do not trust?

Conclusion

- Quick pass through some general evaluation topics which I think are very important
 - Types of evaluation
 - Evaluation is hypothesis testing
 - Common problems in experimental design
 - Replication
 - Challenges in evaluating LLMs