We Should Evaluate Real-World Impact

Ehud Reiter

ehudreiter.com

Who is Ehud?

- Prof of Computing at Aberdeen Uni (UK)
  - » Formerly Chief Scientist of Arria NLG
- Working on NLG, esp evaluation, since the late 1980s
  - » Often on healthcare applications
- NLG blog: ehudreiter.com
- Book: *Natural Language Generation*
  - » https://ehudreiter.com/book/

Evaluating Impact

- When are NLG/NLP systems actually useful in real-world contexts?
  - » We claim LLMs changing world…
- Evaluate usefulness by measuring *impact* when system is deployed
  - » Change in *Key Performance Indicators* (KPI) such as writing time, decision quality, etc
- Mostly ignored by NLP community

Paper

- Much of this talk based on

E Reiter (2025). We should evaluate real-world impact. *Computational Linguistics*

https://doi.org/10.1162/coli.a.18

https://arxiv.org/abs/2507.05973

Example

- Do LLMs help professional software dev?
  - » LLMs do great on coding benchmarks
- Becker et al (2025) eval impact
  - » Randomised controlled trial
  - » Asked devs to do tasks with/without LLMs
  - » Measured productivity (KPI)
  - » Productivity higher *without* LLM assistance

J Becker et al (2025). Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity. https://arxiv.org/abs/2507.09089

Need impact eval

- BM do *not* predict real-world utility
  - » LLM great on BM, poor on productivity
- Subjective opinions do not predict util
  - » Devs thought that LLMs helped them
- Only way to eval utility is to measure it
- Ideally in many contexts
  - » Becker et al looked at specific context (dev, task, LLM tool, etc)
  - » Meta analysis (sys review) across contexts

Impact Eval Matters

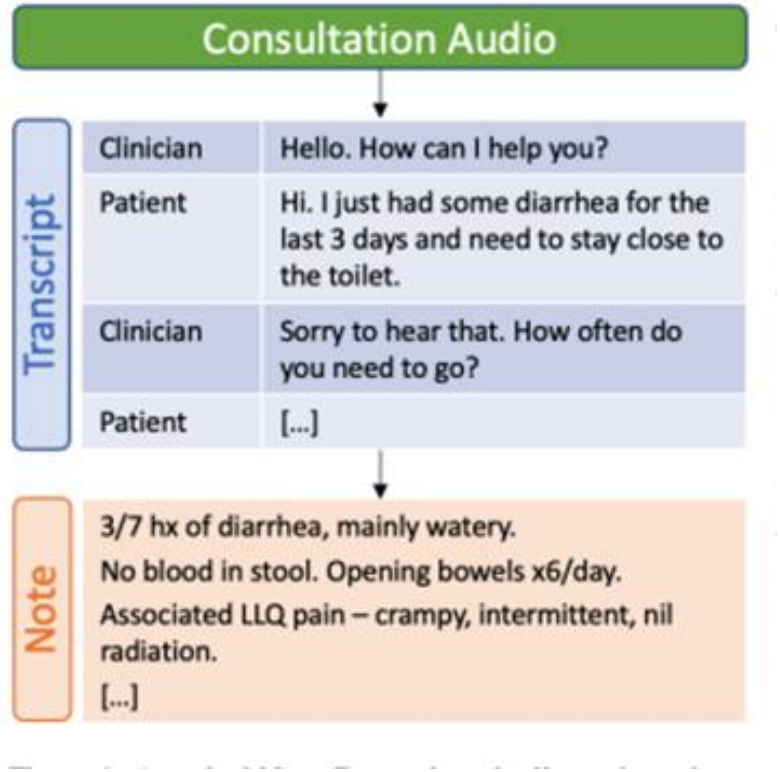
- Important to eval utility/impact!
  - » Users know when to use NLP
  - » Researchers know issues lowering impact

Contents

- *What is impact evaluation*
- Types of impact evaluation
- Survey of impact eval in ACL Anthology
- Discussion
- Encouraging impact evaluation

Example: NoteGenerator

- Summarise doctor-patient consultations for patient record



Moramarco et al (2022). Human Evaluation and Correlation with Automatic Metrics in Consultation Note Generation. ACL-2022

Knoll et al (2022). User-Driven Research of Medical Note Generation Software. NAACL-2022

Moramarco (2024). Evaluation of Medical Note Generation Systems. PhD thesis, Aberdeen

Many types of evaluation

- *Metrics*: Used Bleu, Bertscore, etc to compare NLG texts to corpus text
- *Human intrinsic*: Asked docs for Likert ratings, error annotation of NLG texts
- *Human extrinsic:* Measured post-edit time in experimental context
- ***Impact****: Measured post-edit time in real-world usage*

Real-World Impact

- Deployed system, compared clinician time required to create summary (KPI)
  - » Average time writing summary manually (before system)
  - » Average time post-editing generated sum
- 9% decrease on average
  - » Depends on type of consultation
  - » Questionable return-on-investment (ROI)
- Slightly fewer errors (second KPI)

Impact Evaluation

- Effect of system when deployed in real-world usage (workflows, tasks, users)
- Measure the KPIs that are important to users and stakeholders
- Common and expected in medicine and other fields

Contents

- What is impact evaluation
- *Types of impact evaluation*
- Survey of impact eval in ACL Anthology
- Discussion
- Encouraging impact evaluation

Types of impact evaluation

- Compare NLP system KPIs to baseline
  - » Controlled trial, A/B test, before-and-after study
- Measure KPIs (no baseline)
  - » Observational studies
- Other possibilities (not discussed here)
  - » Qualitative
  - » Longitudinal studies

Controlled Trial

- (*do LLMs help software developers*?)
- Allocate subjects to different groups
  - » Different intervention in each group
  - » Measure average KPI in each group
- Ideally randomised and blinded
- Best evaluation in medicine
  - » Required when eval drug effectivenes

Example: Smoking Cessation

- STOP: NLG sys for smoking cessation
  - » Generate letter based on questionnaire
- Three groups in randomised CT
  - » Received STOP letter, fixed letter, thank you letter
- KPI: Stopped smoking 6 months later
- Outcome: Fixed letter group did best

E Reiter et al (2003). Lessons from a Failure: Generating Tailored Smoking Cessation Letters. *Artificial Intelligence*

A/B testing

- Different users see different versions of a web page or app
  - » Measure average KPI for each version
- Same idea as controlled trial, but used in non-clinical area such as marketing
  - » Eg, which web page design will lead to the highest number of sales?

Example: MT

- Russell and Gillespie (2016) compared two MT systems for translating user reviews on an ecommerce site
- Some visitors saw texts from MTSysA, others from MTSysB
- Measured: pages per visit, items added to cart, items bought (KPIs)

B Russell et al (2016). Measuring the behavioral impact of machine translation quality improvements with A/B testing. *Proc of EMNLP16*

Before-and-After Study

- Measure KPIs before and after an NLP system is introduced
- Ideally based on routinely acquired data, no special experiment needed
- Not ideal that KPIs are measured at different times

Example: NoteGenerator

- Previously mentioned
- KPI: Compare
  - » Time to manually write summaries before NoteGenerator
  - » Time to post-edit NoteGenerator summaries when system is deployed
- Also compared errors (KPI)

Observational Study

- Measure KPIs when people use a system
- No comparison to baseline
  - » May not be possible, eg novel task with no baseline

Example: credit alerts

- Nygaard et al (2024) describe NLP system that alerts credit officers to relevant news
- KPI: How many of these alerts provide new and useful info to credit officers

A Nygaard et al (2024). News Risk Alerting System (NRAS): A Data-Driven LLM Approach to Proactive Credit Risk Monitoring. *Proc of EMNLP24 Industry Track*

Types of impact evaluation

- Controlled (clinical) trial
- A/B test
- Before-and-after study
- Observational study

- Unfortunately all of above are rare in NLP!

Contents

- What is impact evaluation
- Types of impact evaluation
- *Survey of impact eval in ACL Anthology*
- Discussion
- Encouraging impact evaluation

Impact Eval in ACL Anthol

- Question: How many papers in ACL Anthology include impact evaluations?
- Structured survey to investigate

- Answer: 0.1% (1 in a 1000)

Stage 1: EMNLP24 Ind Track

- Checked all 122 papers in EMNLP24 Industry Track
  - » 10 papers (8%) included an impact eval
- 7 of these papers only briefly mentioned impact eval, after describing metric eval in detail
  - » Authors did not regard it as important…

Example of brief desc

After observing significant improvements during offline simulations, we launched an online A/B experiment on live traffic to determine the impact of our proposed approach on geocode learning. We performed the model dial-up in a phased manner -- 10%, 50%, and 100% traffic. We observed statistically significant improvements during one week of dial-up in each phase. During the A/B test period, our approach learnt geocodes for a few hundred thousand shipments, where we observed 14.68% improvement in delivery precision and 8.79% reduction in delivery defects. (*from Maheshwary et al 2024*)

- No details, unclear what really happened
- Metric study (prec/recall) described in detail

S Maheshwary et al (2024). Pretraining and Finetuning Language Models on Geospatial Networks for Accurate Address Matching. *Proc of EMNLP24 Industry Track*

Impact evaluation

- I am far more impressed by reduction in delivery defects than by metric numbers!
- However, authors seem to have opposite opinion, metrics more import than impact

Keywords

- Looked for best title/abstract keywords for identifying EMNLP24 Industry track papers with impact eval
  - » A/B test, deployed (present in 6/10 papers)
- Used these in larger study

Stage 2: Anthology

- Downloaded Anthology bib file (106K papers) on 12 March 2025
- Searched for title/abstract keywords
  - » A/B test, clinical trial, before-and-after evaluation, pre-post evaluation
  - » Also “deployed” (from Stage 1 analysis)
  - » 537 papers identified
- Checked these for impact evaluation

Results

- Found 45 Anthology papers which included impact studies
  - » 0.04% (45/106000) !!
- Probably missed some papers
  - » 4/10 Stage 1 papers did not use these KW
- Best guess is that 0.1% of Anthology papers contain impact eval
  - » Still very low!

Results

- Only 15 (1/3) of these papers described impact eval in moderate detail
  - » At least one column and one table
- A/B test most common technique
- 32 (70%) appeared in Industry Tracks

Impact eval in ACL Anth

- Extremely rare
  - » 0.1% (1 in 1000) of Anthology papers
- Usually secondary to metric study
  - » Impact eval not described in detail
- Mostly in Industry Track

- Disappointing…

Other evidence

- My student Mengxuan Sun found no impact eval in a survey of NLP in cancer care (Sun et al 2025)

M Sun et al (2025). The role of natural language processing in improving cancer care: A scoping review with narrative synthesis. *Artificial Intelligence in Medicine*

Contents

- What is impact evaluation
- Types of impact evaluation
- Survey of impact eval in ACL Anthology
- *Discussion*
- Encouraging impact evaluation

Need some impact eval

- I don't expect all NLP papers to include impact eval!
  - » Not appropriate for theory, math, ling, etc
  - » Also not for work-in-progress
- But should be more than 0.1%
  - » Impact eval should be part of the evaluation ecosystem for NLP
  - » Some eval of real-world utility!

Users want this

- Lack of clinical trials is barrier to deploying NLP in healthcare
  - » From AI enthusiast in UK NHS
- Lawyers, teachers, and regulators (among others) also want impact data

Impact eval pub elsewhere

- Clinical trial evaluations of NLP *are* published in medical literature
- Could say medics should do this eval
- I think better if NLP researchers involved and aware of outcomes
- We claim NLP tech is changing world
  - » Why are we so reluctant to measure real world impact?

Research culture

- NLP culture focus on test set eval
  - » ML, DARPA Common Task
  - » Impact evaluation ignored
- PhD students working on NLP evaluation never heard of impact eval
- Very hard to change culture

Contents

- What is impact evaluation
- Types of impact evaluation
- Survey of impact eval in ACL Anthology
- Discussion
- *Encouraging impact evaluation*

Education

- Make researchers aware of impact eval
- Talks, papers, books, blogs
  - » I am trying…
  - » Great if other people do this as well
- Tutorials and panels at conferences
  - » May suggest this for 2025-26, anyone want to help?

Incentives

- Publication: special themes or tracks for papers with impact evaluation
  - » Reviewers who care about impact
  - » Most ARR reviewers not aware/interested
- Funding: Special calls which require impact evaluation
  - » Perhaps in medicine initially

Final Thoughts

- Impact evaluation is important to people who use NLP
- Important to NLP researchers who want to use tech to help people
- Extremely rare in NLP research
- Lets change this!