Human Evaluation

Ehud Reiter

Human Evaluation

- Get people to evaluate generated text
  - Directly ask them
  - Measure impact of text on tasks (eg, decision-making)
- Focus on basics here
  - Doing a human evaluation well is essential!
  - Poor human evaluation meaningless

Human Evaluation

- The best way to evaluate text-generation systems when done properly
  - Especially when evaluating content and/or utility
  - Especially when looking for subtle problems
- Need high-quality human evaluation to validate metrics

Discussion

- Has anyone done a human evaluation?
- What happened?

Contents

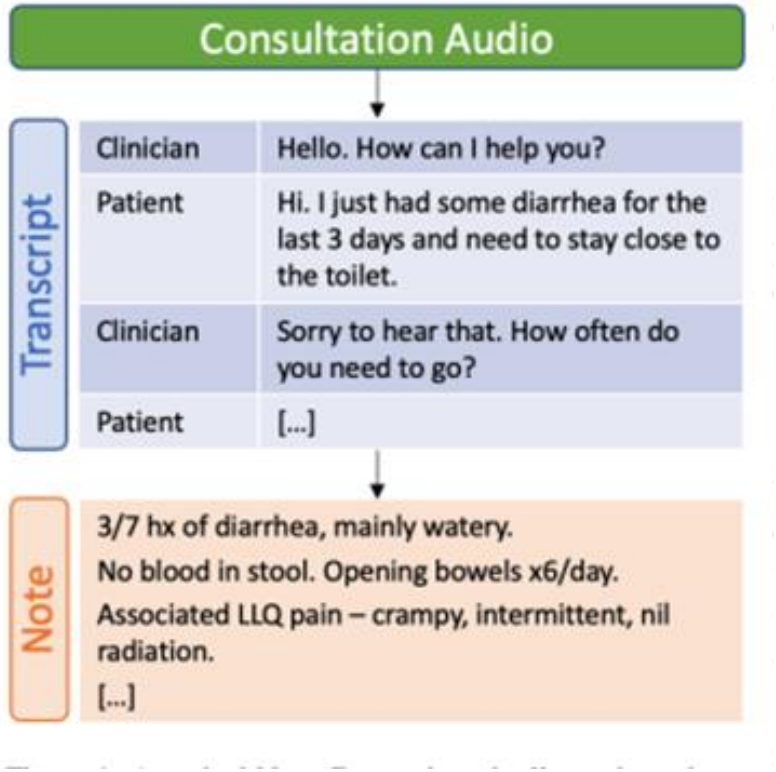
- *Types of human evaluation*
- Examples of good evaluations
- Experimental design and what goes wrong
- Research ethics

Many types of human eval

- Subjective ratings/rankings
  - Likert scales, rank outputs, etc
- Annotating specific problems
  - Mark up hallucinations, errors, other prob
- Task-based
  - Impact on performance, eg decision qual
- Real-world impact
  - Use system for real, measure impact on KPI

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

Ratings

- Asked doctors to use system and give ratings on 1-10 scale
  - How easy is it to write up your notes within the ten minutes allocated for the consultation?
  - Etc
- Subjective opinion

Annotating problems

- Asked doctors to read summaries, find and categorise individual errors
  - Contradiction (*no family history of bowel issues; father has history of colon cancer*).
  - Incorrect statement (statement: *brother has diabetes*. Correction: *mother has diabetes*)
  - Nonsensical (*No recent unwell with diarrhoea*)
  - etc

Task performance

- Timed how long it took doctors to post-edit generated summaries (in lab, not real life)
  - Very important KPI (key performance indic)
  - Depends on UI, user training, workflow as well as quality of generated texts

Impact

- Deployed system, compared
  - Average time writing summary manually (before system)
  - Average time post-editing generated sum
- 9% decrease on average
  - Depends on type of consultation
- Slightly fewer errors

Types of human eval

- Ratings/ranking – easiest to do, but subjective
- Annotating problems – more objective and (in my view) reliable
- Task-based – good, but depends on UI, training, workflow as well as text quality
- Impact – best eval, but hardest to do

Contents

- Types of human evaluation
- *Examples of good evaluations*
- Experimental design and what goes wrong
- Research ethics

Ratings/rankings: Marine weather forecasts

- Goal: evaluate computer-generated marine weather forecasts
  - Also evaluated human (corpus) forecasts
- Subjects: forecast readers (work in offshore oil industry)
- Two experiments
  - Rankings: Showed people two forecasts from same data, ask which is better from perspective of different quality criteria
  - Ratings: Showed people one forecast, asked them to rate quality criteria on a Likert scale
  - Also showed source numerical weather data
- Outcome: NLG forecasts sometimes better than human!

Rank texts

Numerical prediction data

Time	Wind Direction	Wind Speed
0600	S	18
0900	S	19
1200	S	22
1500	SSE	23
1800	S	24
2100	S	22
0000	SSW	20

Text (a)

Wind(10M): S'LY 15-20 BECOMING 22-28 BY THIS EVENING. LATER VEERING S-SW 18-22

Text (b)

Wind(10M): S 16-21 BACKING SSE 21-26 BY MID AFTERNOON, THEN VEERING S BY EARLY EVENING AND SSW 18-23 BY MIDNIGHT

Please cross (or tick) one box for each questions

	(a)	(b)	both same
Which text is easiest to read?	☒	☐	☐
Which text is most accurate?	☐	☒	☐
Which text is most appropriate?	☒	☐	☐

Comments (if any)

For clarity, less text is better. The shift of direction from S to SSE would have minimal operational impact and would probably go un-noticed unless frequent observations were being made.

Rate text

Wind data:

Date/hour	Wind dir	Wind speed
10/06	S	13
10/09	SSW	13
10/12	SW	13
10/15	SW	13
10/18	SSW	12
10/21	S	12
11/00	SSE	14

Wind statement:

S 10-15 VEERING SW AROUND MIDDAY BACKING S DURING THE NIGHT INCREASING BY END OF PERIOD TO 13-18

Your evaluation:

Please score the above wind statement in terms of the following criteria. The higher the score, the better the forecast (1=very bad, 7=very good).

Clarity and readability: ○1 ○2 ○3 ○4 ○5 ○6 ○7

Accuracy and appropriateness: ○1 ○2 ○3 ○4 ○5 ○6 ○7

Submit scores and get next forecast

Explanation of scores:

- 1 = Worse than any forecast you have seen.
- 2 = As bad as very poor forecasts you have seen.
- 3 = Fairly bad, but you have seen worse.
- 4 = Acceptable, neither particularly good nor bad.
- 5 = Fairly good, but you have seen better.
- 6 = As good as very good forecasts you have seen.
- 7 = Better than any forecasts you have seen.

Annotation: sports stories

- Goal: Identify factual errors in computer-generated basketball summaries
  - Specific errors and error types (not just counts)
- Subjects: Mechanical Turkers
  - Vetted, knew basketball, decently paid, worked with us for years
  - Motivated to do a good job because they liked the work
- Process: Marked up errors using MS Word (familiar)
  - Also categorised errors
  - 3 people annotated each text, used majority opinion
- Outcome: Better understanding of hallucinations

Team & Player Data

TEAM	W	L	H1-PTS	H2-PTS	PTS	FG%
Grizzlies	5	0	46	56	102	.486
Suns	3	2	52	39	91	.559

Player	TEAM	PTS	REB	AST	BLK	STL
Marc Gasol	Grizzlies	18	5	6	0	4
Isaiah Thomas	Suns	15	1	2	0	1

Mistakes in text (annotated)

The Memphis Grizzlies (5-**2**) defeated the Phoenix Suns (3-2) **Monday** 102-91 at the **Talking Stick Resort Arena** in Phoenix. The Grizzlies had a **strong** first half where they **out-scored** the Suns **59**-**42**. Marc Gasol scored 18 points, **leading** the Grizzlies. **Isaiah Thomas** added 15 points, he is averaging **19** points on the season so far.

Mistake categories

Name	Player, Team, day of week, etc.
Number	Number, in any form.
Word	Word or phrase that is not **Name/Number**.
Context	Something that is contextually wrong.
Not Checkable	Impossible/time-consuming to check.
Other	Any other error.

Task-based: Decision support for doctors

- Goal: Determine where NLG summaries of patient data helped doctors and nurses make better decisions
- Subjects: doctors and nurses
- Process
  - Used data from 5 years previously
  - Subjects say either visualization, NLG text, or doctor-written summary of relevant patient data
  - Asked doctors to recommend intervention
  - Evaluated whether intervention was correct
- Outcome: doctor texts best, NLG texts similar to visuals

Partial example

BACKGROUND

Born at 26 weeks + 4 days gestation, birth weight 800 grams, he is now 2 weeks old.

He was on CPAP but yesterday was re-intubated because of more frequent apnoeas and bradycardias. Ventilator settings are CMV, rate 35, pressures 18 / 4, iT 0.3 seconds and 35% oxygen. He is in an incubator set at 33ºC. Treatment includes vancomycin, netilmicin, caffeine, and a platelet transfusion. He is pink, active and responsive to handling.

There have been numerous desaturations to the 70s and the inspired oxygen has been adjusted in response to these; the most recent change was an increase from 29 to 35% at 14:09.

Impact: Smoking cessation

- Goal: Determine if NLG smoking-cessation advice helped
- Subjects: smokers
- Process
  - Smokers filled out questionnaire about their smoking
  - Received either (A) NLG text, (B) fixed text, (C), simple thank-you letter
  - Waited 6 months, asked if stopped smoking (verified with saliva sample)
  - Compared cessation rates in groups
- Outcome: fixed text as good as or better than NLG texts
  - Negative result!

Questionnaire (extract)

SMOKING QUESTIONNAIRE

Please answer by marking the most appropriate box for each question like this: ☒

Q1 Have you smoked a cigarette in the last week, even a puff?

YES ☒ — Please complete the following questions

NO ☐ — Please return the questionnaire unanswered in the envelope provided. Thank you.

Please read the questions carefully. If you are not sure how to answer, just give the best answer you can.

Q2 **Home situation**:

Live alone ☐ | Live with husband/wife/partner ☒ | Live with other adults ☐ | Live with children ☒

Q3 **Number of children** under 16 living at home …… boys ……1……. girls

Q4 **Does anyone else in your household smoke?** *(If so, please mark all boxes which apply)*

husband/wife/partner ☒ other family member ☒ others ☐

Q5 **How long have you smoked for?** …10… years

Tick here if you have smoked for less than a year ☐

Letter (extract)

Smoking Information for Heather Stewart

You have good reasons to stop...

People stop smoking when they really want to stop. It is encouraging that you have many good reasons for stopping. The scales show the good and bad things about smoking for you. They are tipped in your favour.

You could do it...

Most people who really want to stop eventually succeed. In fact, 10 million people in Britain have stopped smoking - and stayed stopped - in the last 15 years. Many of them found it much easier than they expected.

Although you don't feel confident that you would be able to stop if you were to try, you have several things in your favour.

- You have stopped before for more than a month.
- You have good reasons for stopping smoking.
- You expect support from your family, your friends, and your workmates.

We know that all of these make it more likely that you will be able to stop. Most people who stop smoking for good have more than one attempt.

Overcoming your barriers to stopping...

You said in your questionnaire that you might find it difficult to stop because smoking helps you cope with *stress*. Many people think that cigarettes help them cope with stress. However, taking a cigarette only makes you feel better for a short while. Most ex-smokers feel calmer and more in control than they did when they were smoking. There are some ideas about coping with stress on the back page of this leaflet.

You also said that you might find it difficult to stop because you would *put on weight*. A few people do put on some weight. If you did stop smoking, your appetite would improve and you would taste your food much better. Because of this it would be wise to plan in advance so that you're not reaching for the biscuit tin all the time. Remember that putting on weight is an overeating problem, not a no-smoking one. You can tackle it later with diet and exercise.

And finally...

We hope this letter will help you feel more confident about giving up cigarettes. If you have a go, you have a real chance of succeeding.

With best wishes,

The Health Centre.

Contents

- Types of human evaluation
- Examples of good evaluations
- *Experimental design and what goes wrong*
- Research ethics

Experimental design for human eval

- *Research question*: Which quality criteria are we trying to estimate, what baselines (if any) do we compare to
  - Depends on use case, research goals
- *Type:* What type of evaluation
  - Ratings/rankings are cheapest, easiest, quickest
    - I personally prefer ranking, but other people prefer ratings
    - Not great with near-human quality LLM texts
  - Annotation more expensive but more meaningful
    - My default
  - Task-based great for utility, but contextual (eg, UI matters)
  - Impact best but hardest to do
    - Ethical challenges

Experimental design: subjects

- Who will do experiment, how many, how recruited
  - Essential to get this right!
- Who are they, how recruited
  - Crowdworkers: Mechanical Turk, Prolific, Upwork, etc (*easiest)*
  - Colleagues, friends, family, students (*cheapest*)
  - Hire specific individuals to do experiment (*best for specialized work*)
- Filter out inappropriate subjects
  - Vetting process as well as recruitment filters
- Numbers: ideally determined by statistical power calculation
  - Depends on experiment (happy to discuss)

Subjects: What goes wrong

- Subjects do not take task seriously, just click at random
  - Or use ChatGPT to respond
  - Big danger with remote crowdworkers
    - Vet and monitor!
  - Less of an issue with other types of subjects and in-person experiments
- Not enough subjects
  - Many NLG evaluations are “under-powered” in statistical sense
  - More subjects is always better
- Subjects not representative of real users
  - Don’t ask CS students to evaluate medical NLG system

Experimental design: material

- Which texts should subjects evaluate?
- Numbers are usually much smaller than in automatic eval
  - Choose texts which are useful for evaluation
- How choose
  - Random subset of test data corpus (*default*)
  - Divide test data into categories, randomly choose within categories
    - Gives more insight into performance if done well
  - Explicitly add difficult edge/boundary cases (*unusual in academics*)

Material: What goes wrong

- Data contamination
- Selected texts are all common/straightforward cases
  - Especially if randomly selected
- Cannot cover all edge/boundary cases
  - 100s/1000s of these!
  - Major issue in safety

Experimental design: Procedure

- What will subjects do in detail
  - Instructions, UI, training, attention checks, etc
- Which texts does each subject see
  - If each subject just sees some of the texts
  - Should be balanced; Latin Square can be useful
- How is data recorded
  - What happens if network outage, subject gets sick, etc
- Specify procedure *before* experiment starts
  - Don’t make it up as you go along!
  - Doing a small pilot first can help

Procedure: What goes wrong

- Based on ReproHum survey
- Insufficient guidelines and training
  - Subjects don’t understand what they are supposed to do
  - See J Ruan et al (2024). https://aclanthology.org/2024.naacl-long.441/
- Poor UI confuses subjects
- Insufficient monitoring/vetting of subjects
  - Subjects click randomly and this isn’t detected
- Buggy experimental code
- Task too complex and subjective

Experimental Design: Analysis

- How is experiment data analysed?
- Statistical analysis should always be done
- Calculate inter-annotator agreement if possible (eg, Kappa)
  - Do subjects agree with each other?
  - If not, experiment may not be reliable
- Qualitative analysis highly recommended
  - I always let subjects make free-text comments about experiment
- Outliers may be removed if indicate data not reliable

Analysis: What goes wrong

- Bugs in analysis code
- Aggressive outlier policy distorts results
- Post-hoc hypotheses not clearly identified

Example: Basketball NLG

- Experimental design for system that generates summaries of BasketBall matches

Contents

- Types of human evaluation
- Examples of good evaluations
- Experimental design and what goes wrong
- *Research ethics*

Research ethics

Experiments with human subjects must be ethical!

- Subjects give informed consent
  - Bad: don’t tell subjects they are in experiment (Facebook)
- Subjects not coerced into being in experiment
  - Bad: “Strongly encourage” students to be subjects
- Subjects and third parties not harmed
  - Bad: Upset people by showing them racist language
- Personal/sensitive data is protected
  - Bad: publish or leak medical data, including smoking history

Ethical approval

- Most universities in Western countries have ethics boards
  - In Aberdeen, any experiment which involves human subjects must be approved by a university research ethics board
    - Takes weeks if straightforward
  - US ethics rules tighter, because of past problems
  - China doesn’t do this yet, but it is coming
- Rules for companies are weaker
  - Lawsuits may be more of an issue than formal approval
- Overall trend is for stronger ethical rules and governance

Big issue for impact evaluation

- If deploying system for real, must show it will not harm people!
  - Note generator: wont lead to poor care because of errors
  - Smoking cessation: wont upset smokers
  - Etc
- Challenging
  - Can argue that benefits outweigh risks (as in medicine)
  - But then approval takes a long time

Conclusion

- Human evaluation is the best way to evaluate many qual criteria
  - Especially accuracy and utility
- Many ways of doing human eval
  - Ranking/ratings are most common, but may be least meaningful
- Experiments must be done well!
  - Otherwise results mean nothing