From Metrics to Meaning: Learning Qualitative Research Methods as a Community

Adarsa Sivaprasad, Yujun Wang, Mengxuan Sun, David M. Howcroft, Ehud Reiter
Department of Computer Science, University of Aberdeen, Scotland, UK
{a.sivaprasad.22,y.wang2.22,m.sun.22,david.howcroft,ehud.reiter}@abdn.ac.uk

Abstract

The field of natural language generation (NLG) is naturally quite quantitative in nature and mindset. However, quantitative methodologies can give us measures but not tell us what to measure or why it matters. Here we document our efforts to conduct a qualitative study of changing evaluation practices in NLG-adjacent research communities. We share our experience designing a semi-structured interview procedure, the kinds of conversations that result, how they differ from our expectations as NLP researchers, and where we are going next in our study, along with early insights from the initial coding of interview transcripts.

1 Introduction

Natural language generation (NLG) evaluation has received sustained and growing attention, resulting in the publication of metascientific papers surveying existing practice (Howcroft et al., 2020; Schmidtova et al., 2024, i.a.) and recommending best practices (van der Lee et al., 2021; Gehrmann et al., 2023, i.a.). Such work provides an important overview of which practices are commonly adopted, how well they are executed, and what recommendations researchers can follow. However, beyond learning an overview of current practices, research also requires critical reflection on the strengths and limitations of studies and systems. Such reflection can reveal not only what works, but also where and why current approaches may fail, thereby helping researchers derive more informative conclusions from their results.

Error analysis is an example of such reflection and is severely under-reported in Natural Language Generation systems (van Miltenburg et al., 2021, i.a.). Researchers strongly agree that error analyses are useful and want to see more, but barriers remain (van Miltenburg et al., 2023, i.a.). The authors liken error analysis to "eating broccoli": people know it's healthy, but not everyone enjoys the taste or finishes the plate. They consequently propose measures to lower these barriers and encourage more systematic and transparent error reporting in the NLG literature. More broadly, learning from quantitative studies might help us measure the problems or tasks we already know. However, quantitative studies cannot tell us about problems that have not yet been reported. They also tell us little about how different researchers view the field, why they make certain choices, what values and trade-offs shape those choices, and how their experiences influence the practices we see in published papers.

Addressing this gap requires a qualitative approach that speaks with researchers and practitioners about their experiences and how they interpret the changes the field has undergone. This paper presents a metascientific perspective: sharing our experience conducting a qualitative study as researchers from primarily quantitative backgrounds to make the process more accessible to other quantitatively-minded researchers looking to understand the *why* and *how* questions relevant to their research. We also share initial insights from our first phase of interviews and analysis.

2 Why qualitative methods?

The overall aim of our project is to investigate how evaluation practices in NLG are currently understood, applied, and interpreted by researchers. In particular, we want to know how experienced NLP researchers understand, justify, and critique evaluation practices, and what lessons from these experiences can inform guidance for future evaluation.

While prior surveys and meta-analyses document what evaluation practices appear in the literature (Howcroft et al., 2020; Reiter, 2018; van der Lee et al., 2021; Gehrmann et al., 2023; Reiter, 2025), they cannot explain why these practices persist, how community norms around them formed, or why well-documented limitations do not lead

Proceedings of the 19th International Natural Language Generation Conference: Short Papers, pages 74–83
October 17-21, 2026 ©2026 Association for Computational Linguistics

to changes in practice. Answering these questions requires access to the reasoning, justifications, and professional judgments of researchers who make evaluation decisions, something the literature alone cannot provide.

This is precisely the class of research question for which qualitative methods are the established instrument across scientific fields. In medicine, Pope and Mays (1995) established the purpose of qualitative methods on the grounds that they enable us to access areas not amenable to quantitative research, such as lay and professional beliefs. Further, qualitative description is a prerequisite for good quantitative research, particularly in areas with little prior investigation. They help uncover institutional pressures and tacit norms (McDonald et al., 2005; Bierbaum et al., 2020). The parallel to NLG evaluation is direct: the community continued using BLEU even when it was known to be a poor proxy for generation quality.

We adopt a sequential mixed-methods design. In the first phase, semi-structured interviews with senior NLP researchers provide the depth and flexibility needed to elicit reasoning in participants' own terms, enabling systematic comparison across participants while remaining open to unexpected lines of argument. The group of authors iteratively analyse the data through rounds of discussion (Braun and Clarke, 2006). The later sections document the analysis process in detail. The planned second phase is a community survey derived from the interview-identified themes that extends these findings to a broader and more diverse population, including early-career researchers and participants from a wider range of geographic and institutional contexts. This converts the expert-identified patterns into quantifiable community-level positions.

Recent work has begun to apply qualitative approaches to natural language processing (NLP) as a field (Gururaja et al., 2023). Our study departs from the anonymised design of prior work: participants were given the option to be named, and many consented, enabling us to connect individual positions to specific intellectual trajectories and give readers the contextual grounding to interpret each account. This is, to our knowledge, the first qualitative interview study of NLG evaluation practice to combine expert testimony with named attribution. We present it here as a methodological contribution demonstrating how qualitative approaches can be applied rigorously in NLP metascience.

3 Interview design

The study design reflects the study's core methodological commitment to eliciting causal reasoning rather than reporting behaviour. Semi-structured interviews are an appropriate instrument for this purpose. A structured interview or survey requires pre-specifying the categories that matter, thereby foreclosing unexpected reasoning or retrospective narration. A fully unstructured approach would sacrifice comparability across participants. The semi-structured interview, with a blend of closed- and open-ended questions, each accompanied by a follow-up question, balances thematic consistency across interviews with the flexibility to follow a participant's argument where it leads (Kallio et al., 2016). Participants can articulate their reasoning in their own terms, construct retrospective accounts of how current norms emerged, and defend or complicate positions under gentle probe, producing evidence that no fixed-response instrument can generate (Adeoye-Olatunde and Olenik, 2021).

The interview team developed the interview guide through discussion, with guidance from the senior author, and tested it in a pilot interview. Interviews were organised around three thematic areas, moving deliberately from past to present to future: (1) participants' experiences with NLG evaluation practice in their early career, (2) perceived strengths and limitations of current approaches and (3) directions for future NLG evaluation. This temporal arc was designed to drive causal reasoning, prompting participants to connect historical patterns to present conditions and project forward from that diagnosis. We refined the interview questions through three rounds of meetings and online comments. The individual questions followed a similar structure: for each key issue, we asked participants about their views, the reasons behind those views, whether and how past experiences shaped them, and what recommendations they would offer. The full interview protocol is provided in the Appendix A.

This study was approved by the University of Aberdeen Ethics Board (reference number 12852547). All participants provided informed consent prior to the interview, including explicit consent to recording. Participants were given the explicit choice of named or anonymous attribution at the outset. Participants were also offered the opportunity to review their transcript prior to analysis. Table 1 includes some relevant participant details.

PhD Year	Country	Affiliation
2007	Netherlands	Academia
2004	UK	Academia
1990	UK	Academia
2019	Netherlands	Academia
1991	Netherlands	Academia
2013	USA	Industry
2018	USA	Academia
2010	Germany	Industry
2016	Germany	Academia
2009	China	Academia

Table 1: PhD year (used as a proxy for career stage), current country of affiliation, and affiliation for our interviewees

Recruitment was facilitated by the recent workshop (RetroEval, 2026) which brought together researchers and practitioners with a shared interest in NLG evaluation from academia and industry, enabling 6 in-person interviews during the event. We recorded the interviews on a university-managed device and transcribed them using *Microsoft Word*'s transcribe service. We interviewed an additional 4 participants online via *Microsoft Teams*. The products were chosen from a limited set of university-approved tools and worked well for our needs. A member of the research team manually verified transcriptions before sharing them with interviewees for review.

4 Analysis and early insights

While this paper focuses on process rather than analysis, our initial coding has already revealed interesting insights, which we share here after a brief discussion of our coding procedure.

Our approach to coding Coding was conducted in *NVivo*. We planned to follow a primarily deductive approach to coding the transcripts, based on the structure imposed by the questions asked in our interviews. Each of the four author-annotators initially coded one transcript in common along with the transcripts for their own interviewees. Because most of us have not previously conducted this kind of qualitative analysis, we met after this initial attempt at coding to discuss what was (not) working and how to proceed. The *semi*-structured interviews had some structure based on our questions, but not every topic fit neatly into our initial codes, as discussed further in Section 5.

One author coded the full dataset, while the other three authors each coded one-third of the data, with one overlapping interview included to support cross-coder comparison. We initially considered using inter-annotator agreement to assess coding consistency. However, as the analysis progressed, we found that participants' diverse experiences often required interpretation beyond a fixed coding scheme. We therefore adopted an iterative and collaborative thematic analysis approach. All coders could introduce new codes when the existing codebook did not capture relevant concepts, and we updated the codebook iteratively. The coding team then discussed disagreements, overlapping codes, and newly introduced codes, and refined or merged them through group discussion. In these discussions, the team reviewed newly introduced top-level codes with the interview excerpts assigned to them. Where a broader category was derived, one coder implemented the agreed grouping in *NVivo*. We assigned separate subcodes to distinct opinions retaining different perspectives, rather than treating them as contradictions. We grouped related codes into broader categories based on conceptual similarity and their relevance to the research questions. Unifying them coherently with respect to our deductive codes, we arrived at a set of 188 codes, with 17 top-level and 100 second-level labels[1].

Early evaluation practices One thing that became clear from our interviews is that homogeneity as a field has both costs and benefits. They shared that NLP in the '90s did not typically involve what we would recognise today as an evaluation, with performance often not assessed or, when it was, assessed on checklists of specific linguistic constructions (Reiter, 2026). This approach suited researchers exploring a variety of NLG tasks, but community convergence around referring expression generation (REG), formalised by (Dale and Reiter, 1995), enabled the creation of the TUNA shared task on REG (Gatt and Belz, 2009) and a resulting shared standard of evaluation for a limited set of tasks.

Contemporaneous with this convergence was the emergence of BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and other word-overlap metrics. These automated evaluation metrics were appealing because of their ease of use and seeming comparability[2], leading to their rapid adoption despite early calls for caution (Belz and Reiter, 2006) and ongoing criticism in response to their continued widespread use (Novikova et al., 2017).

[1] See Appendix B for an overview.

[2] though see (Post, 2018)

Problematic current practices This persistence of metrics which are not informative or have known weaknesses was especially concerning to our participants, with one saying *"I'm not an anti-metric person in principle, but I think there are good metrics and there are not so good metrics. And there are some metrics that we know are terrible in the sense that they don't really tell us anything, and people are still reporting their results on those things."* The persistent overemphasis on overlap-based metrics in NLG evaluation is extensively documented (Schmidtova et al., 2024, i.a.), yet the reasons for this persistence are less often examined directly.

Our participants' reflections offer a practitioner account of how this situation came about, with one recurring theme being conformity and pressure to keep reporting what others have reported. *"[P]eople are very afraid that if they don't report that they might be negatively reviewed"* and *"conformity pays off"*, but there isn't a commensurate emphasis on *"understanding what those metrics mean"* or *"evaluation metrics that really do take the users more into consideration"*. Participants do acknowledge that *"there are certain real benefits for the community in using [the] same corpora and methods"* but also that *"there are certainly career benefits to doing that"*.

However, the persistence of individual metrics also depends on the ecosystem of available metrics and training. One participant commented that *"[a] lot of people just don't know how to do qualitative evaluations or impact evaluations"* and highlighted that a past proposed tutorial on impact evaluation was rejected, perhaps because the topic was too "niche". However, our interviewees expressed a desire for a greater diversity of evaluation methods and metrics with an emphasis on the real-world impact of our systems, with multiple participants suggesting that researchers need to apply critical thinking to design problem-specific evaluations for their research questions, *"why they're evaluating what they're evaluating, what is [it] specifically [they] hope to learn through their evaluation"*.

5 Challenges in qualitative methods

Qualitative methods bring their own challenges, some of which have already been mentioned.

For one, there is tension between the deductive frame of analysis that follows interview structure and the emergent content of those interviews. Qualitative interviews are not rigid questionnaires: they are guided conversations in which researchers pursue planned topics while remaining responsive to meanings, emphases, and connections introduced by participants (Kvale and Brinkmann, 2015; Mitchell et al., 2012). We designed the interview structure chronologically to elicit causal reasoning from our participants. In practice, participants did not follow this arc cleanly, with causal accounts appearing throughout.

This also required a shift in how we treated labels during analysis, motivating us to introduce cross-temporal codes (e.g., *factors influencing evaluation practice*) rather than limiting ourselves to period-specific observations. In a quantitative annotation study, changing labels or introducing new categories after seeing the data is usually problematic, because labels function as fixed measurement categories. However, freezing the coding scheme during analysis in our present study would exclude many valuable insights from our participants — exactly the insights we aim to capture with a qualitative approach! This kind of movement between deductive and inductive analysis is well established in qualitative research: some approaches begin with theory- or question-driven categories, others derive categories more directly from the data, and hybrid approaches combine both (Hsieh and Shannon, 2005; Braun and Clarke, 2006). For our purposes, the shift from a strictly deductive approach to a more inductive approach for analysis reflects our research goals.

Another issue with interview methods, contrary to our training as quantitative researchers who use machine learning methods, is the small sample size. Interviews and thematic analysis are researcher-time-intensive methods, which limit the number of interviews we can reasonably incorporate (Guest et al., 2006; Malterud et al., 2016). However, with just ten participants, it is not clear whether a theme appearing in three transcripts reflects the community as a whole or these individuals. This is why we plan to use a mixed-methods approach, building on these interviews to design a survey for the wider NLG community. The interviews provide insights into the kinds of questions we should ask and candidate answers for multiple-selection questions grounded in more than our personal knowledge, giving us a stronger foundation for a survey that can provide further insights at a scale that allows us to draw conclusions about the community as a whole.

6 Discussion and conclusion

Our preliminary analyses already suggest that the overemphasis on automatic metrics is best understood not as a knowledge gap but as a coordination failure: a feedback loop in which the dominance of simple metrics became increasingly difficult to break, not because researchers believed in them, but because the individual cost of deviation – a negative review, an incomparable result, an unfamiliar method – is structurally higher than the collective cost of continuing. Further, a lack of training in qualitative and impact-based methods means that even researchers who wish to go beyond standard approaches may lack the methodological grounding to do so credibly.

These preliminary findings could not be found through bibliometric methods or other more quantitative approaches to metascientific inquiry; rather, they are a unique product of the qualitative approach to understanding the problem. In this work-in-progress report, we share our experience designing this qualitative study, some of the challenges encountered along the way that may be especially disconcerting to the quantitative researcher, and argue for the importance of this kind of work.

We hope this report encourages others to apply qualitative methods—not only for metascience but also to understand what to build for whom and why, and what kinds of evaluations actually matter. Our methods would transfer well to interviews of other experts (e.g., medical professionals, lawyers) whose needs and desires or motivations need to be understood for your research questions or to identify an appropriate problem to work on. We encourage you to plan carefully but also to be ready to adapt to the opportunities provided by your participants—this effort to adapt while trying to answer *how* and *why* questions will help you ensure that, when you return to quantifiable measures, you are measuring what matters.

For us, the plan is to iterate on our analysis of these interviews, including more inductive codes to reflect not just the content that fit our interview structure but the themes that were important to our participants. Building on these interviews and the final analyses, we will launch a survey to gather broader perspectives on evaluation practices from a wider group of researchers and practitioners in NLG, with the ultimate goal of providing recommendations for both individual and community action to improve the state of evaluation in our field.

Limitations

The main limitations of this work at present are discussed in Section 5. For most of the author-annotators involved in this work, this is our first interview study using thematic analysis with inductive coding, so we are still learning. We hope that being open about these limitations and sharing our work in progress will make qualitative studies less intimidating for others coming from a more quantitative background.

Ethical considerations

University of Aberdeen Ethics Application Number 12852547. All researchers involved completed the research ethics and governance and research integrity certification. Participants were informed that their participation is entirely voluntary and that they may withdraw from the study at any time without providing a reason. For interview participants who wished to remain anonymous, this was maintained in the analysis and reporting of results.

References

Omolola A Adeoye-Olatunde and Nicole L Olenik. 2021. Research and scholarly methods: Semi-structured interviews. *Journal of the american college of clinical pharmacy*, 4(10):1358–1367.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 313–320, Trento, Italy. Association for Computational Linguistics.

Mia Bierbaum, Jeffrey Braithwaite, Gaston Arnolda, Geoffrey P Delaney, Winston Liauw, Richard Kefford, Yvonne Tran, Bróna Nic Giolla Easpaig, and Frances Rapport. 2020. Clinicians' attitudes to oncology clinical practice guidelines and the barriers and facilitators to adherence: a mixed methods study protocol. *BMJ open*, 10(3):e035448.

Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.

Albert Gatt and Anja Belz. 2009. Introducing shared tasks to nlg: The tuna shared task evaluation challenges. In *Conference of the European Association for Computational Linguistics*, pages 264–293. Springer.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 77:103–166.

Greg Guest, Arwen Bunce, and Laura Johnson. 2006. How many interviews are enough? an experiment with data saturation and variability. *Field methods*, 18(1):59–82.

Sireesh Gururaja, Amanda Bertsch, Clara Na, David Widder, and Emma Strubell. 2023. To build our future, we must know our past: Contextualizing paradigm shifts in natural language processing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13310–13325, Singapore. Association for Computational Linguistics.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Hsiu-Fang Hsieh and Sarah E Shannon. 2005. Three approaches to qualitative content analysis. *Qualitative health research*, 15(9):1277–1288.

Hanna Kallio, Anna-Maija Pietilä, Martin Johnson, and Mari Kangasniemi. 2016. Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. *Journal of advanced nursing*, 72(12):2954–2965.

Steinar Kvale and Svend Brinkmann. 2015. *Interviews*. Sage.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Kirsti Malterud, Volkert Dirk Siersma, and Ann Dorrit Guassora. 2016. Sample size in qualitative interview studies: guided by information power. *Qualitative health research*, 26(13):1753–1760.

Ruth McDonald, Justin Waring, Stephen Harrison, Kieran Walshe, and Ruth Boaden. 2005. Rules and guidelines in clinical practice: a qualitative study in operating theatres of doctors' and nurses' views. *BMJ Quality & Safety*, 14(4):290–294.

Elizabeth D Mitchell, Greg Rubin, and Una Macleod. 2012. Understanding diagnosis of lung cancer in primary care: qualitative synthesis of significant event audit reports. *The British Journal of General Practice*, 63(606):e37.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Catherine Pope and Nick Mays. 1995. Qualitative research: reaching the parts other methods cannot reach: an introduction to qualitative methods in health and health services research. *bmj*, 311(6996):42–45.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ehud Reiter. 2018. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401.

Ehud Reiter. 2025. We should evaluate real-world impact. *Computational Linguistics*, pages 1–13.

Ehud Reiter. 2026. NLG evaluation: Past, present, future. In *Proceedings of the 1st Symposium on Natural Language Generation Evaluations*, pages 8–15, Aberdeen, United Kingdom. Association for Computational Linguistics.

RetroEval. 2026. Retroeval 2026: Symposium on natural language generation evaluations. Accessed: 2026-09-01.

Patricia Schmidtova, Saad Mahamood, Simone Balloccu, Ondrej Dusek, Albert Gatt, Dimitra Gkatzia, David M. Howcroft, Ondrej Platek, and Adarsa Sivaprasad. 2024. Automatic metrics in natural language generation: A survey of current evaluation practices. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 557–583, Tokyo, Japan. Association for Computational Linguistics.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

Emiel van Miltenburg, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Emma Manning, Stephanie Schoch, Craig Thomson, and 1 others. 2021. Underreporting of errors in nlg output, and what to do about it. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 140–153.

Emiel van Miltenburg, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Stephanie Schoch, Craig Thomson, and Luou Wen. 2023. Barriers and enabling factors for error analysis in nlg research. *Northern European Journal of Language Technology*, 9.

A Interview questions

Interview Protocol

Fixed interview questions are limited to facilitate short (10-20 minute) interviews during the RetroEval symposium. Online interviews may be longer because they have fewer time constraints, but we will aim to keep them under 30 minutes so they do not differ too substantially in character from the in-person interviews.
Methodology: In-person interviews will be recorded using the Transcribe feature in Microsoft Word online and stored in University storage. Online interviews will be conducted using Microsoft Teams with automatic transcriptions and stored in University storage.
Transcripts will be sent to participants before analysis as a courtesy, to allow correction or elaboration, and recordings will be deleted once transcripts are finalised.
Depending on participant consent, transcripts may be published with attribution or pseudonymously.
The following describes the Topics to be discussed in the interviews, the goal of this portion of the interview, and the questions to be used as the basis for the interview.

A.1 Interview Duration

- In-person: 10–20 minutes
- Online: Up to 30 minutes

A.2 Interviewer Preparation

Before beginning:

- Confirm participant has received the Participant Information Sheet.
- Confirm consent has been obtained according to study procedures, e.g. the online form https://forms.office.com/e/HV3NagADQR
- Ensure recording/transcription software is active.
- Note participant identifier and interview date.

A.3 Opening Script (2 minutes)

Thank you for agreeing to participate in this study.
As outlined in the participant information sheet, this research aims to understand what lessons researchers and practitioners draw from the evolution of Natural Language Processing evaluation practices and how these lessons shape their expectations and recommendations for the field's future. This interview will take approximately 10 to 20 minutes. With your permission, I will record and transcribe the interview. The transcript will be shared with you before analysis so that you can make corrections or add clarifications if you wish. Once the transcript is finalised, we will delete the recording.
There are no right or wrong answers. We are interested in your experiences, reflections, and perspectives. The questions are framed around 3 topics

- What Early Evaluation Practice Can Teach Us.
- What Current Practice Gets Right and Wrong.
- What research directions and evaluation practices would be relevant in the future NLP landscape

Before we begin, do you have any questions about the study or the interview process?
Do I have your permission to begin recording and start the interview?
[Start Recording]

A.4 Interview Questions

1. To start, how would you describe your field of study?
2. We start with the topic of What Early Evaluation Practice Can Teach Us. The goal is to use your early-career experiences to identify what shaped the field and whether those assumptions should be carried forward or left behind.
    - What would you describe as your early career years in NLP, and what evaluations did you do then?
    - What evaluations were common at the conferences you attended or in the venues where you published?
    - Were any of those particularly noteworthy or exceptional?
3. Moving on to what Current Practice Gets Right and Wrong, we aim to understand the current evaluation landscape as a basis for practical guidance.
    - What best practices do you currently follow in conducting evaluation?
    - Do researchers generally do this?
    - What practices does the community over- and underemphasise? Is there an evaluation practice or quality attribute that is over- and underemphasised?
    - When people do not use best practices, why do you think that is?
    - In your field, what are the expected practices? Does this influence the quality attribute?
    - Are ethical attributes considered as a quality criterion?
    - Is the research ethical? Are IRBs consulted? What measures are taken in involving human participants?
4. Moving to the last part of the study, we want to understand what research directions and evaluation practices would be relevant in the future NLP landscape.
    - If you were advising a PhD student starting an NLP project today, what would you tell them to do differently in evaluation than what the field's current norms would lead them to do?
    - What do you think evaluation practice should look like in ten years? What would future researchers need to do now to move it in that direction?

A.5 Closing Script

Thank you very much for your time and insights.
As mentioned earlier, we will send you a copy of the transcript before analysis. You will have the opportunity to correct errors, clarify points, or add any further comments if you wish.
If you have any questions after the interview, please feel free to contact the research team.
Thank you again for participating.
[Stop recording.]

B Codebook Overview

Top-level code	Representative subcodes
Study field	NLG; linguistics; fairness and bias evaluation; hallucinations; healthcare applications.
Past	Reliance on automatic evaluation over human evaluation; task-specific evaluation; maintaining human oversight when using LLM evaluators.
Current practices	Hypothesis-driven evaluation; over-emphasis on automatic metrics and SOTA benchmarks; under-emphasis on data leakage, disaggregated evaluation, and real-world evaluation; reliability of LLM evaluators.
Future directions	Critical thinking about evaluation; real-world problems and applications; meta-evaluation; diversity of evaluation methods; consistent terminology; qualitative and impact evaluation; integrating evaluation earlier in system development.
Factors influencing evaluation choices	Peer-review pressure; citation incentives; publication pressure; pursuit of SOTA performance; resource constraints; preference for low-risk publishable research; evaluation as secondary to system development; selective reporting.
Ethical considerations	Diverse perspectives; growing attention to ethics; IRB considerations; burdensome ethics processes; limited influence of ethical concerns within organisations; limited integration of ethical criteria in computer science evaluation.

Table 2: Overview of the high-level codebook. Representative subcodes are illustrative, not exhaustive.