Last Words

We Should Evaluate Real-World Impact

Ehud Reiter

University of Aberdeen, UK
e.reiter@abdn.ac.uk

The ACL community has very little interest in evaluating the real-world impact of NLP systems. A structured survey of the ACL Anthology shows that perhaps 0.1% of its papers contain such evaluations; furthermore most papers that include impact evaluations present them very sketchily and instead focus on metric evaluations. NLP technology would be more useful and more quickly adopted if we seriously tried to understand and evaluate its real-world impact.

1. Introduction

The medical community places great emphasis on clinical trials that assess the real-world effectiveness of new medications and other interventions; papers in education and engineering also regularly present data on real-world impact of new techniques. But in NLP, very few papers (even in Industry or Applications tracks) present data on real-world effectiveness of deployed systems. We regularly claim that LLMs and other NLP technologies are changing the world (Maslej et al. 2025), but we are reluctant to provide data on how deployed NLP systems improve real-world **key performance indicators** (KPIs); I will refer to such changes in KPIs as *impact* below.

Of course providing data on changes in real-world KPIs is hard, much harder than calculating benchmark and metric scores on a test set, or asking crowdworkers to assess the quality of an output. Measuring impact requires deploying the system in production usage, monitoring its effect on users, and getting permission to publicly release this data. But still, other fields are publishing papers on real-world impact of LLMs, including medicine (Duggan et al. 2025) and software engineering (Pandey et al. 2024), so it can be done.

It is also notable that even when real-world impact is measured, many NLP papers treat this as secondary to a metric evaluation. For example, Maheshwary, Paul, and Sohoney (2024) carefully describe a metric-based evaluation, and then in a single paragraph summarize a real-world A/B impact evaluation that shows significant improvements in important KPIs (Section 4.1). Similarly, Yoon et al. (2024) carefully describe a metric-based evaluation, and then briefly mention that a real-world before-and-after

Action Editor: Saif Mohammad. Submission received: 27 April 2025; revised version received: 19 June 2025; accepted for publication: 22 June 2025.

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

study showed very impressive KPI improvements. In short, many NLP researchers do not seem to think real-world impact is important (at least in an academic paper), even when they have data on this.

In this article I first discuss and give examples of impact evaluation. I then show in a literature survey that such papers are rare in NLP, and conclude with a discussion and suggestions for encouraging more NLP researchers to evaluate real-world impact.

2. Impact Evaluation

There are many ways of evaluating NLP systems. Most evaluations involve running several systems (including a baseline) on a *test set* of data or scenarios, and evaluating how well the NLP systems do on the test set. Such evaluations are most commonly done using automatic metrics such as precision and recall for classification, BLEU score (Papineni et al. 2002) for text generation, and (more recently) using LLMs as judges (Zheng et al. 2023). Some studies use human evaluation; the most common form of this is asking human participants to rate outputs or annotate problems in outputs (Reiter 2025). Human evaluation can also be *extrinsic* (Jones and Galliers 1995), which involves measuring the effect the NLP system has on helping users do something, such as make good clinical decisions (Portet et al. 2009).

Such evaluations have been very successful in driving the development of core NLP technologies in areas such as speech recognition and machine translation (Liberman and Wayne 2020). But they do not give complete understanding of how well NLP systems work in complex real-world applications. Also, researchers may optimize test set performance in ways that do not actually increase real-world utility; this is an aspect of Goodhart's Law[1] ("When a measure becomes a target, it ceases to be a good measure").

An impact evaluation, in contrast, involves deploying a system and measuring changes in KPIs in real-world usage; since we are directly measuring the KPIs that we care about, Goodhart's Law is not a problem. Note that while an impact evaluation can assess multiple systems, different systems usually interact with different users. So there is usually *not* a single test set which is presented to all the systems being evaluated.

Two papers by Moramarco and colleagues illustrate the difference between extrinsic evaluation on a test set and evaluating impact. Moramarco et al. (2022) present an extrinsic evaluation of systems that summarized clinician-patient consultations. In this evaluation, they used a test set of mock consultations (consultations between a clinician and an actor) (Papadopoulos Korfiatis et al. 2022). They asked clinicians to listen to the mock consultations, read computer-generated summaries from several different models (and also a human-written corpus text), and post-edit the summaries to make them accurate and acceptable; they measured the time needed to post-edit. Moramarco (2024) did an impact evaluation of a single system in this domain. In this study Moramarco measured post-edit time when the final system was deployed and used with real patients in actual consultations (Section 3.3); doctors post-edited consultation summaries during or immediately after the consultation, as part of their normal clinical workflow.

3. Examples of Impact Evaluation Techniques

Impact can be evaluated by comparing KPIs between people who use an NLP system and people who are in control group(s); techniques for doing this include clinical trials,

[1] https://en.wikipedia.org/wiki/Goodhart%27s_law.

A/B testing, and before-and-after studies. Impact can also be evaluated in observational studies where KPIs are measured just in users of the NLP system. I give examples below.

Note that impact can also be evaluated *qualitatively*, using user observations and feedback, expert analyses of case studies, error analysis, example outputs, etc. While the NLP community strongly emphasizes quantitative results, Kapoor, Henderson, and Narayanan (2024) argue that in an AI in law context, good qualitative studies are more meaningful than quantitative studies. Qualitative observations can also be very important in building good theoretical models. Regardless, since qualitative analyses of impact in NLP are usually supplements to quantitative analyses, I will focus on quantitative analyses in this article.

3.1 Clinical Trials

Clinical trials are popular in medicine. In such experiments, patients are allocated to different groups, which receive different treatments; KPIs are measured and compared between groups. There are many types of clinical trials (Greenhalgh 1997); the most highly regarded are randomized (patients are randomly allocated to groups) and blind (patients do not know which group they are in).

The STOP project (Reiter et al. 2001) built and evaluated a Natural Language Generation system that generated letters giving advice on smoking cessation, using data from a smoking questionnaire that participants filled out. The evaluation was a randomized controlled clinical trial:

- We recruited 2,553 smokers, and randomly allocated them to three groups of roughly equal size: STOP, FixedLetter, ThankYou.
- All smokers filled out the smoking questionnaire.
- Smokers in the STOP group were sent STOP letters, smokers in the FixedLetter group were sent a fixed letter (manually edited version of STOP's default letter), and smokers in the ThankYou group were just sent a letter thanking them for being in our study.
- After six months, we asked smokers if they had stopped smoking (this was our KPI); we verified this with a saliva test for nicotine residues.

Note that this was not a blind trial, since patients knew which group they were in.

Unfortunately, the results showed that STOP was not effective; cessation rates were in fact highest in the FixedLetter group, although the difference was not statistically significant.

For more information on the STOP experiment, see Lennox et al. (2001) and Reiter, Robertson, and Osman (2003).

3.2 A/B Testing

A/B testing is a common technique used for evaluating Web page and resources. Users coming to the Web page are randomly directed to either a baseline or new page, and their behavior is monitored. In a sense, A/B testing adapts clinical trial methodology to non-clinical use cases, including sales and marketing.

Russell and Gillespie (2016) used A/B testing to evaluate whether a customized machine translation (MT) system did better than a generic MT system when translating user reviews on an e-commerce Web site. The evaluation was structured as follows:

- 88,106 visitors to an e-commerce Web site who needed translations were randomly allocated to GenericMT or CustomMT groups.
- Visitors in the GenericMT group saw the output of the generic MT system; visitors in the CustomMT group were shown the output of the custom MT system.
- Researchers measured three KPIs: pages per visit, whether items were added to a cart, and whether items were bought (conversion rate).

Results showed that the CustomMT group had (statistically significant) higher values for all KPIs; for example, conversion rate increased by 8.7%.

3.3 Before-and-After Studies

A *before-and-after* study (sometimes called *pre-post study*) measures changes in KPIs after a new tool is introduced. This means that KPIs based on the new tool are measured after pre-tool KPIs, which is not ideal (its better to measure both at the same time). However, such analyses have the advantage that they may not require a special experiment to be organized; they can instead be based on data collected during routine operations.

Unfortunately, I am not aware of any paper in the ACL Anthology that reports before-and-after studies at an acceptable level of detail. A before-and-after study is described in detail in Chapter 7 of Moramarco's Ph.D. thesis (Moramarco 2024). This evaluated the impact of a tool that generated summaries of doctor-patient consultations (which are needed for the patient record). The evaluation was structured as follows:

1. 20 clinicians were selected, and the time they spent writing summaries in real consultations (before the tool was deployed) was calculated.
2. The summarization system was deployed, and the time these clinicians spent editing the computer-generated summaries was calculated, again in real consultations. This was done over a 7-month period. Comparing this to (1) measured changes in the KPI of time spent creating the summary.
3. 20 manually written summaries and 20 edited-computer-summaries (from the same clinicians) were carefully checked for errors. This measured changes in the KPI of number of mistakes in summaries.

Results were positive but not overwhelmingly so. Post-editing computer summaries was 9% faster than manually writing summaries. This is relatively small, so the tool may not be cost-effective from a Return on Investment (*ROI*) perspective. There were also slightly fewer mistakes in the post-edited summaries.

3.4 Observational Studies

The techniques described above all compare an NLP system to something else, and report changes in KPIs. But if an NLP system is doing a novel task that has not been done before, then comparison is difficult and the paper may just report observed KPIs without comparisons.

For example, Nygaard et al. (2024) describe an NLP system for credit officers that alerts them to relevant news about companies that have borrowed money. The system was evaluated as follows:

- It was deployed for two types of alerts, sentiment and mergers/acquisitions.
- 3,500 alerts produced by the system were manually classified as new information, recently considered (already known), irrelevant, etc.

Twenty-three percent of the alerts were new information, and hence useful to the credit officers.

4. Literature Search

In order to get a better understanding of impact evaluations in NLP venues, I performed two structured literature searches of the ACL Anthology. The first was on papers in the EMNLP 2024 Industry Track; this gave me a better understanding of keywords that identified papers that reported real-world impact. I then used these keywords to search the entire ACL Anthology.

I focused on papers in the ACL Anthology because my goal is to understand the NLP community's perspective on impact evaluation. The medical literature contains papers that describe clinical trials of NLP systems (e.g., Meystre and Haug 2008) and before-and-after studies of NLP systems (e.g., Duggan et al. 2025), but my focus is the NLP literature, not the medical one.

4.1 Papers in the EMNLP 2024 Industry Track

Industry Tracks at xACL conferences solicit papers about real-world applications, so seem especially likely to include impact studies. The largest Industry Track at the time of writing was at EMNLP 2024, which included 122 papers. I read the abstracts of all 122 papers; if the abstract suggested any kind of impact evaluation, I read the paper body. I excluded papers that described technical details of deployment but did not give impact data, such as Singhal et al. (2024).

This resulted in 10 papers (8% of total) that gave impact data; most of these used A/B testing. Most of these papers gave very short descriptions of the impact study. For example, Maheshwary, Paul, and Sohoney (2024) describe a metric study in detail, and then describe an A/B impact study in a few sentences as follows:

> *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*

Table 1
Impact papers in EMNLP 2024 Industry Track. *Detailed* means description of impact study in the paper was at least 0.5 pages and included at least one figure or table.

Type	Count	Detailed
clinical trial	0	–
A/B test	6	2 (33%)
before-after	1	0 (0%)
observational	3	1 (33%)
TOTAL	10	3 (30%)

> *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.*

Since many details are missing, it is difficult to interpret the impressive-sounding improvements in KPIs.

Because of this problem, I checked which papers (A) devoted at least half a page (one column) to the impact study and (B) included at least one table or figure from the study. Only three of the ten papers met this criteria. See Table 1 for more information.

I was also disappointed that *none* of the papers gave the kind of detailed information about the impact evaluation that is present in the above-mentioned medical papers (Meystre and Haug 2008; Duggan et al. 2025). These papers are shorter than full-length xACL papers, so the problem is not size limitations.

Anyways, one goal of this exercise was to identify keywords (for title and abstract) that could be used to search the rest of the Anthology for papers that included an impact evaluation. Unfortunately, terminology varied widely in the Industry Track papers. The best keywords I found were "A/B test" and "deployed," but these only matched six of the ten papers.

In short, this analysis suggests that (at least in Industry Track papers):

- The most common type of impact evaluation in NLP is A/B testing, and the second most common is observational studies.
- When an impact evaluation is presented in an NLP paper, it is usually secondary to a metric evaluation, and often described very briefly. This is the case even in Industry Track papers written by people from companies.
- The best title/abstract keywords for identifying papers with impact evaluations are "A/B test" and "deployed," but many papers do not mention either of these keywords in their title or abstract.

Incidentally, EMNLP 2024 as a whole included around 3,000 papers (main conference, findings, workshops, Industry Track). I am only aware of one EMNLP paper not in the Industry Track that presented impact studies (Dai et al. 2024), so overall 0.37% (11/3,000) of EMNLP 2024 papers included impact evaluation.

4.2 Papers in ACL Anthology

I downloaded the ACL Anthology bib file on 12 March 2025; it contained 105,850 papers. I used Zotero to search for papers whose title or abstract contained "deployed," "A/B test," "clinical trial," "before-and-after evaluation," or "pre-post evaluation." This resulted in 537 papers.

I checked these 537 papers and identified 41 papers that gave quantitative impact data from a deployed system. As mentioned above, I excluded papers that gave purely qualitative data, such as Varges et al. (2012).

So all together, if I include the 4 EMNLP 2024 Industry Track papers that did not use any of the above keywords, I found 45 papers in the ACL Anthology that included impact studies; this is 0.04% of the papers in the Anthology.

Of course there may be additional papers in the Anthology that report impact but do not use any of the above keywords. The stats from Section 4.1 (40% of the ten papers did not use any of the keywords) suggest that the actual percentage of Anthology papers that include impact studies may be closer to 0.1%. This is an estimate, but my expectation is that the percentage of Anthology papers that include impact evaluation is between 0.05% and 0.2%, with 0.1% being my best estimate.

From a venue perspective:

- 32 (71%) of the papers appeared in Industry Tracks.
- 7 (16%) of the papers appeared in xACL conferences (including EMNLP, COLING, and Findings) outside of Industry Tracks.
- None of the papers appeared in journals.
- 6 (13%) of the papers appeared in other Anthology venues (mostly workshops).

Tables 2 and 3 break down the 45 papers by study type and keyword. Overall, this is similar to the results in Section 4.1:

- Very few Anthology papers include impact evaluations.
- A/B tests are the most common type of impact evaluation, followed by observational studies.

Table 2
Number of impact papers found in ACL Anthology, by study type. *Detailed* means description of impact study in the paper was at least 0.5 pages and included at least one figure or table.

Type	Count	Detailed
clinical trial	1	1 (100%)
A/B test	37	10 (27%)
before-after	1	0 (0%)
observational	6	3 (50%)
TOTAL	45	14 (31%)

Table 3
Number of impact papers found in ACL Anthology, by abstract/title keyword. Some papers match multiple keywords, so TOTAL is less than the sum of per-keyword counts. *Detailed* means the description of the impact study in the paper is at least 0.5 pages and includes at least one figure or table.

Keyword	**Count**	**Detailed**
clinical trial	1	1 (100%)
A/B test	22	7 (32%)
before-after	0	–
deployed	24	6 (25%)
pre-post evaluation	0	—
(EMNLP Industry Track)	10	3 (30%)
TOTAL	45	14 (31%)

Table 4
Impact papers found in ACL Anthology, where the description of the impact study in the paper was at least 0.5 pages and included at least one figure or table.

Paper	**Type**	**Application Area**
Reiter et al. (2001)	clinical trial	smoking cessation
Russell and Gillespie (2016)	A/B	machine translation
Elsafty, Riedl, and Biemann (2018)	A/B	recommender system
Chapman et al. (2020)	observational	identifying COVID cases
Srivastava et al. (2021)	A/B	customer service
Liao and Fares (2021)	observational	customer service
Joshi et al. (2022)	A/B	multi-label classification
Golobokov et al. (2022)	A/B	generating advertisements
Liu et al. (2022)	A/B	relevance matching
Rubin et al. (2023)	A/B	virtual assistant
Dai et al. (2024)	A/B	machine translation
Mohankumar et al. (2024)	A/B	selecting bids for spons search
Chen et al. (2024)	observational	identifying important searches
Nygaard et al. (2024)	observational	credit risk monitoring

- Only a third of papers that report impact studies describe them in even moderate detail.

Table 4 lists the 14 papers I found in the Anthology that report impact studies in at least moderate detail.

5. Discussion

The above literature survey suggests that papers that report real-world impact are extremely rare in the ACL Anthology, and, even when they do appear, the impact study is often described very briefly and is treated as a supplement to a metric study. I personally find a reduction in real-world delivery defects to be far more impressive than precision/recall/accuracy metrics, but clearly Maheshwary, Paul, and Sohoney (2024) have the opposite opinion (Section 4.1). In short, the ACL community seems to have little interest in evaluating real-world impact of NLP systems.

Note that impact evaluation is related to *ethical evaluation*, since the underlying purpose of ethical evaluations is to understand whether NLP systems can harm people or society. This is best measured by deploying systems and measuring harm-related KPIs, since harmful behavior often occurs when NLP systems are used in complex real-world contexts that their developers did not anticipate. From this perspective, impact evaluation has links to attempts to understand ethical considerations, such as Mohammad (2022) and Karamolegkou et al. (2025).

5.1 When Is Impact Evaluation Appropriate?

I do not expect all NLP papers to include an impact evaluation! In medicine, impact evaluations are mainly used in papers that are intended to guide real-world decision-making; they are not used in theoretical, modeling, and speculative papers, for example. Similarly in NLP, I expect to see impact evaluations (at least some of the time) in papers that propose NLP models or systems for real-world usage; these papers are often (but not always) part of Industry or Applications tracks. Impact evaluation is not appropriate for papers that are not applied and focus on theoretical issues, modeling fundamentals, cognitive science, etc. I also do not expect to see impact evaluation in speculative, work-in-progress, or position papers.

Perhaps most fundamentally, impact evaluation should be part of the NLP research "ecosystem" and happen at least in some cases; this will provide NLP as a whole with feedback on the real-world utility of different approaches.

5.2 Barriers to Impact Evaluation

There are barriers to impact evaluation in NLP. One is simply that many NLP researchers have little knowledge of impact evaluation; this is discussed in Section 6. Practicalities can also be a barrier, including duration and cost of evaluation, and in many cases the need to include a partner who can deploy a system and measure its impact. These problems are less of an issue for companies, which may be why 71% of the impact evaluations I found were in Industry Track papers.

Perhaps more fundamentally, the NLP's community lack of knowledge and interest in real-world impact evaluation may reflect its culture and the strong influence of machine learning (and perhaps the DARPA Common Task Method [Liberman and Wayne 2020]), which focus on evaluations based on test sets. Impact evaluations usually do not evaluate multiple systems on the same test set (Section 2), since this is very difficult to do when evaluating real-world usage of systems; hence they do not fit the test-set evaluation model.

5.3 Users Want Impact Evaluation

Communities that use NLP technology are frustrated by the lack of impact evaluation. For example, Hiebel et al.'s (2025) review of NLG in clinical contexts states that `High scores on output quality metrics do not necessarily imply that deploying the system will benefit users.` This point was also made very strongly to me by someone involved in assessing AI systems for use in part of the UK

National Health Service; the lack of proper effectiveness evaluations (ideally randomized controlled clinical trials) is a major barrier in deploying AI systems in the NHS.

In the legal community, Kapoor, Henderson, and Narayanan (2024) state that `the kinds of legal applications we can legitimately use AI for should be determined by the evaluations that reflect these uses of AI in the real world.` I am not aware of *any* studies of real-world impact of NLP in legal contexts that have been published in the ACL Anthology. Lawyers may be more willing than doctors to use NLP without strong evaluation, but they are concerned that they will lose their license if they use misbehaving AI.[2] Solid real-world evaluation will increase adoption and will also provide crucial feedback to NLP researchers working on AI in law.

Concerns have also been expressed in other communities. For example, in education, Ganesh et al. (2023) point out that the performance of an educational dialogue system is much worse in real classrooms.

Regulators are also aware of the need to test and evaluate AI systems in real-world usage. Article 60 of the EU Artificial Intelligence Act[3] waives many of the Act's restrictions when AI systems are being tested in real-world usage.

In short, it is important to evaluate the real-world impact of NLP solutions, and this is acknowledged by both users and regulators. Of course, the NLP community could leave this task to non-NLP researchers. However, if we truly want to develop useful technology, it is much better if we are involved in the evaluations, and also if results from these evaluations are fed back to the NLP research community.

6. Encouraging More Impact Evaluation

I would love to see more impact evaluations in NLP papers! Achieving this will not be easy, since it requires changing the research culture of NLP to focus more on real-world impact and less on SOTA-chasing, and changing research culture is difficult and also slow. However, I do think we can make progress via better *education* and *incentives*.

Education: Many NLP researchers and Ph.D. students have told me that they do not know how to evaluate impact; they were not taught this and have not previously thought about it. I hope that this article will help raise awareness of impact evaluation, including examples of how it is done; I also discuss impact evaluation in a recent book on NLG (Reiter 2025). It would be very useful to have more education and awareness raising about impact evaluation, perhaps via tutorials, panel discussions, and/or keynotes at major NLP conferences.

Incentives: Researchers of course respond to incentives, and publication venues and/or funders could provide incentives to encourage impact evaluation. For example, conferences and other venues could have special tracks or themes about impact, perhaps initially within Industry Tracks. Likewise funders could have special programs or calls that require impact evaluations, perhaps initially in domains such as medicine where this is especially important.

7. Conclusion

If we want NLP technology to be used and genuinely help people, we need to understand its real-world impact on KPIs in deployed production usage. Insights from this

[2] `https://www.cbsnews.com/colorado/news/colorado-lawyer-artificial-intelligence-suspension/`.

[3] `https://artificialintelligenceact.eu/article/60/`.

will help us develop more useful technology; in some fields (especially medicine), they are also an essential prerequisite to large-scale adoption.

Unfortunately the ACL community currently shows minimal interest in evaluating real-world impact. I do not expect every paper in the Anthology to include an impact evaluation, but the number should be more than 0.1%! It is also frustrating that even when impact studies are done for commercial reasons, they are usually treated as being secondary to metric evaluations and also are usually not properly described.

Our community has developed amazing technology that has great potential to change the world; serious evaluation of its real-world impact will make NLP technology more useful and encourage more people to use it.

Acknowledgments

My thanks to Liesl Osman and Scott Lennox for introducing me to impact evaluation and clinical trials, back in the 1990s. Also thanks to my students and colleagues over the years who helped carry out impact-related evaluations, and to the many people who have discussed impact evaluation with me. I would also like to thank Simone Balloccu, Daniel Braun, Nava Tintarev, and the members of the current Aberdeen NLP group for their very useful feedback about this article, and the reviewers for their many helpful suggestions.

References

Chapman, Alec, Kelly Peterson, Augie Turano, Tamára Box, Katherine Wallace, and Makoto Jones. 2020. A natural language processing system for national COVID-19 surveillance in the US Department of Veterans Affairs. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, 7 pages.

Chen, Zhiyu, Jason Ingyu Choi, Besnik Fetahu, and Shervin Malmasi. 2024. Identifying high consideration E-commerce search queries. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 563–572. https://doi.org/10.18653/v1/2024.emnlp-industry.42

Dai, Huangyu, Ben Chen, Kaidi Chen, Ying Han, Zihan Liang, and Wen Jiang. 2024. Contrastive token learning with similarity decay for repetition suppression in machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3247–3261. https://doi.org/10.18653/v1/2024.findings-emnlp.185

Duggan, Matthew J., Julietta Gervase, Anna Schoenbaum, William Hanson, III Howell, John T., Michael Sheinberg, and Kevin B. Johnson. 2025. Clinician experiences with ambient scribe technology to assist with documentation burden and efficiency. *JAMA Network Open*, 8(2):e2460637–e2460637. https://doi.org/10.1001/jamanetworkopen.2024.60637, PubMed: 39969880

Elsafty, Ahmed, Martin Riedl, and Chris Biemann. 2018. Document-based recommender system for job postings using dense representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 216–224. https://doi.org/10.18653/v1/N18-3027

Ganesh, Ananya, Jie Cao, E. Margaret Perkoff, Rosy Southwell, Martha Palmer, and Katharina Kann. 2023. Mind the gap between the application track and the real world. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1833–1842. https://doi.org/10.18653/v1/2023.acl-short.156

Golobokov, Konstantin, Junyi Chai, Victor Ye Dong, Mandy Gu, Bingyu Chi, Jie Cao, Yulan Yan, and Yi Liu. 2022. DeepGen: Diverse search ad generation and real-time customization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 191–199. https://doi.org/10.18653/v1/2022.emnlp-demos.19

Greenhalgh, Trisha. 1997. How to read a paper: Assessing the methodological quality of published papers. *BMJ*, 315(7103):305–308. https://doi.org/10.1136/bmj.315.7103.305, PubMed: 9274555

Hiebel, Nicolas, Olivier Ferret, Karën Fort, and Aurélie Névéol. 2025. Clinical text generation: Are we there yet? *Annual Review of Biomedical Data Science*, 8:173–198. https://doi.org/10.1146/annurev-biodatasci-103123-095202, PubMed: 40101215

Jones, Karen Sparck and Julia R. Galliers. 1995. *Evaluating Natural Language Processing Systems: An Analysis and Review*. Springer.

Joshi, Ashutosh, Shankar Vishwanath, Choon Teo, Vaclav Petricek, Vishy Vishwanathan, Rahul Bhagat, and Jonathan May. 2022. Augmenting training data for massive semantic matching models in low-traffic E-commerce stores. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 160–167. https://doi.org/10.18653/v1/2022.naacl-industry.19

Kapoor, Sayash, Peter Henderson, and Arvind Narayanan. 2024. Promises and pitfalls of artificial intelligence for legal applications. *Journal of Cross-disciplinary Research in Computational Law*, 2(2). https://doi.org/10.2139/ssrn.4695412

Karamolegkou, Antonia, Sandrine Schiller Hansen, Ariadni Christopoulou, Filippos Stamatiou, Anne Lauscher, and Anders Søgaard. 2025. Ethical concern identification in NLP: A corpus of ACL Anthology ethics statements. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11618–11635. https://doi.org/10.18653/v1/2025.naacl-long.580

Lennox, A. Scott, Liesl M. Osman, Ehud Reiter, Roma Robertson, James Friend, Ian McCann, Diane Skatun, and Peter T. Donnan. 2001. Cost effectiveness of computer tailored and non-tailored smoking cessation letters in general practice: Randomised controlled trial. *BMJ*, 322(7299):1396. https://doi.org/10.1136/bmj.322.7299.1396, PubMed: 11397745

Liao, Ling Yen and Tarec Fares. 2021. A practical 2-step approach to assist enterprise question-answering live chat. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 457–468. https://doi.org/10.18653/v1/2021.sigdial-1.48

Liberman, Mark and Charles Wayne. 2020. Human language technology. *AI Magazine*, 41(2):22–35. https://doi.org/10.1609/aimag.v41i2.5297

Liu, Ziyang, Chaokun Wang, Hao Feng, Lingfei Wu, and Liqun Yang. 2022. Knowledge distillation based contextual relevance matching for E-commerce product search. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 63–76. https://doi.org/10.18653/v1/2022.emnlp-industry.5

Maheshwary, Saket, Arpan Paul, and Saurabh Sohoney. 2024. Pretraining and finetuning language models on geospatial networks for accurate address matching. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 763–773. https://doi.org/10.18653/v1/2024.emnlp-industry.58

Maslej, Nestor, et al. 2025. The AI index 2025 annual report. https://hai.stanford.edu/ai-index/2025-ai-index-report

Meystre, Stephane M. and Peter J. Haug. 2008. Randomized controlled trial of an automated problem list with improved sensitivity. *International Journal of Medical Informatics*, 77(9):602–612. https://doi.org/10.1016/j.ijmedinf.2007.12.001, PubMed: 18280787

Mohammad, Saif. 2022. Ethics sheets for AI tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8368–8379. https://doi.org/10.18653/v1/2022.acl-long.573

Mohankumar, Akash Kumar, Gururaj K, Gagan Madan, and Amit Singh. 2024. Improving retrieval in sponsored search by leveraging query context signals. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1489–1498. https://doi.org/10.18653/v1/2024.emnlp-industry.109

Moramarco, Francesco. 2024. *Evaluation of Medical Note Generation Systems*. Ph.D. thesis, University of Aberdeen.

Moramarco, Francesco, Alex Papadopoulos Korfiatis, Mark Perera, Damir Juric, Jack Flann, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. Human evaluation and correlation with automatic metrics in consultation note generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5739–5754. https://doi.org/10.18653/v1/2022.acl-long.394

Nygaard, Adil, Ashish Upadhyay, Lauren Hinkle, Xenia Skotti, Joe Halliwell, Ian C. Brown, and Glen Noronha. 2024. News risk alerting system (NRAS): A data-driven LLM approach to proactive credit risk monitoring. In *Proceedings of the*

2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 429–439. `https://doi.org/10.18653/v1/2024.emnlp-industry.32`

Pandey, Ruchika, Prabhat Singh, Raymond Wei, and Shaila Shankar. 2024. Transforming software development: Evaluating the efficiency and challenges of GitHub Copilot in real-world projects. *arXiv preprint arXiv:2406.17910*.

Papadopoulos Korfiatis, Alex, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. PriMock57: A dataset of primary care mock consultations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 588–598. `https://doi.org/10.18653/v1/2022.acl-short.65`

Papineni, Kishore, 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. `https://doi.org/10.3115/1073083.1073135`

Portet, Francois, Ehud Reiter, Albert Gatt, Jim Hunter, Somayajulu Sripada, Yvonne Freer, and Cindy Sykes. 2009. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, 173(7):789–816. `https://doi.org/10.1016/j.artint.2008.12.002`

Reiter, Ehud. 2025. *Natural Language Generation*. Springer. `https://doi.org/10.1007/978-3-031-68582-8`

Reiter, Ehud, Roma Robertson, A. Scott Lennox, and Liesl Osman. 2001. Using a randomised controlled clinical trial to evaluate an NLG system. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 442–449. `https://doi.org/10.3115/1073012.1073069`

Reiter, Ehud, Roma Robertson, and Liesl M. Osman. 2003. Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, 144(1):41–58. `https://doi.org/10.1016/S0004-3702(02)00370-3`

Rubin, Jonathan, Jason Crowley, George Leung, Morteza Ziyadi, and Maria Minakova. 2023. Entity contrastive learning in a large-scale virtual assistant system. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 159–171. `https://doi.org/10.18653/v1/2023.acl-industry.17`

Russell, Ben and Duncan Gillespie. 2016. Measuring the behavioral impact of machine translation quality improvements with A/B testing. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2295–2299. `https://doi.org/10.18653/v1/D16-1251`

Singhal, Bhavuk, Anshu Aditya, Lokesh Todwal, Shubham Jain, and Debashis Mukherjee. 2024. GeoIndia: A Seq2Seq geocoding approach for Indian addresses. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 395–407. `https://doi.org/10.18653/v1/2024.emnlp-industry.29`

Srivastava, Manisha, Yichao Lu, Riley Peschon, and Chenyang Li. 2021. Pretrain-finetune based training of task-oriented dialogue systems in a real-world setting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 34–40. `https://doi.org/10.18653/v1/2021.naacl-industry.5`

Varges, Sebastian, Heike Bieler, Manfred Stede, Lukas C. Faulstich, Kristin Irsig, and Malik Atalla. 2012. SemScribe: Natural language generation for medical reports. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2674–2681.

Yoon, Chang Oh, Wonbeen Lee, Seokhwan Jang, Kyuwon Choi, Minsung Jung, and Daewoo Choi. 2024. Language, OCR, form independent (LOFI) pipeline for industrial document information extraction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1056–1067. `https://doi.org/10.18653/v1/2024.emnlp-industry.79`

Zheng, Lianmin, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, 36:46595–46623.