

Conference Report: COLM 2026*

Roman Klinger

2026-10-23

Introduction

It's been a while since I wrote a conference report. My [last one](#) was about [ACL 2025](#), but I went to two more conferences afterwards: [EACL 2026](#), which was an interesting experience because of its comparably small size (in contrast to ACL or EMNLP these days), but also because of the place (Rabat, Morocco), which I did personally not like for reasons of personal safety and censorship by the hosting country, that a couple of people I talked to experienced. Further, I've been to [LREC 2026](#) in Palma de Mallorca, a conference that is not considered a flagship venue by some people in the CL/NLP community, but the place to be to report on data sets and resources. It's also a very welcoming venue for less fashionable topics and good discussions.

Nevertheless, I didn't find the motivation to write a conference report about these two conferences, because in the end, they were not very surprising: A lot of LLM work, many interesting papers, both well organized, but nothing I felt that needed to be pointed out in particular. Still, maybe I'll summarize my thoughts about the "smaller" venues in CL/NLP at some point in a blog post. Another reason is that I am constantly overwhelmed by the amount of work I am faced with as a still comparably new full professor.

That situation didn't change too much, but the COLM conference I attended is comparably new, doesn't live in the ACL/CL/LREC world as much, and has a different focus. Therefore, I felt it might be more helpful for some people to read my impressions about the conference to decide if they might want to consider submitting to COLM in addition to their regular ACL Rolling Review submission cycle, and/or COLING and LREC, or TACL and the Computational Linguistics Journal.

Let's get started...

The Third Annual Conference on Language Modeling

The [COLM 2026](#) took place beginning of October in San Francisco. It's only the third iteration of the venue, and I remember 2024, when the organizers reached out to the community about it. It's an interesting venue – language models have been around for a very long time. They are generally models that estimate probabilities of language –

*Published as blog post at <https://www.romanklinger.de/blog/2026-10-23-colm/>

given a word sequence, what is the likelihood that it would be found like that in the real world (or a subsample of it which is used to estimate the model). Another definition could be more focused on prediction and generation, following formal grammars [Chomsky \(1956\)](#), but I want to focus on the statistical perspective here. Such probabilities for a sequence of words can be estimated for instance by counting the frequency of particular n-grams (groups of consecutive words of length n), nicely described in [Jurafsky and Martin \(2026\)](#). While these models have been around for decades, language models entered the world of many people only with the visibility of neural language models and their use in chat systems which are considered to be helpful, such as [Gemini](#) or [ChatGPT](#). These systems are not language models, but they use language models as part of their system.

As these models are sometimes considered general problem solvers (though this idea is [much older](#)), they also and particularly solve language tasks such as generation and understanding. Therefore, they dominate the CL/NLP conferences these days; and these conferences increase in size, and my impression is that non-LM topics are pushed to the side, at least to some degree. This is not a problem in COLM, of course: COLM is designed to focus on language models, and to be honest, this gave me quite some peace of mind. I didn't need to search for other methods I like around CL/NLP (like probabilistic graphical models, for instance), and there was no need to be disappointed. I first thought that this conference would feel narrow. It did not. Not at all.

Review Process and Statistics

The conference was still different from *ACL/LREC/COLING conferences. The review process took place via [OpenReview](#), the platform that ACL moved to after softconf went down a couple of times on the paper submission deadline day, presumably due to large submission numbers. OpenReview does, however, offer more possibilities than just reviewing – it has the idea of being open, and also offers to publish the papers afterwards. Anyway, the differences between [ARR](#) (the rolling review process happening for ACL and COLING these days) and COLM felt minor. The program chairs mentioned that hierarchies in COLM are more shallow than in ACL (reviewers+AC+PC, but no senior area chairs in COLM), but as a reviewer and author, I didn't perceive this too much. The reviews we got for our papers were of very good quality.

The conference received 2939 submissions and accepted 856 papers. These 29% are similar to the acceptance rates of main papers in ACL conferences together with Findings papers, which are often presented as well. I did not perceive a qualitative difference in the program of COLM to ACL or EMNLP. In fact, I have only seen one paper where I wondered how it slipped through in the review process - I typically think that a bit more often (but that could also be because I am more knowledgeable in ACL topics than language model topics).

Interesting was the distribution of countries. While ACL in Vienna had 19% authors from the US and 50% authors from China (Germany had 3%), the COLM distribution was different:

- USA: 62%
- China: 10%
- Canada: 3.9%

- Great Britain: 3.3%
- South Korea: 3.2%
- Germany: 2.6%
- Switzerland: 2%

That is of course not a surprise. Chinese scholars more often have issues traveling to the US than to Europe, and of course there are practical reasons that San Francisco is easier to reach from other places in the US.

It was, however, the first conference which I attended which actually sold out: authors needed to buy their tickets earlier, but then, only 3000 people were admitted altogether. Another difference was that (I only heard that from somebody, so I am not sure if it's accurate) 1/3 of the participants were not from academia, a number which was claimed to be higher in other CL/NLP venues (again, not sure if true).



Figure 1: Poster Session

Organization

The conference was organized very differently than other (big) conferences I attended:

- It was a single track conference, except for the workshop days, there were never multiple oral sessions taking place in parallel.
- The conference was structured in 6 slots (3 days, morning/afternoon), where each slot started with a keynote or a panel, followed by 4 spotlight presentations (which were also available in the poster session), followed by the poster session.

That was easy to follow, but there was no chance to choose between oral presentations, when one session wasn't too much in the own interest. Nevertheless, it felt less stressful than other conferences, because no opportunity/chance/need to run between sessions.

The poster sessions were not thematically structured, as much as I could see. That was also not the case at ACL in Vienna, and I find that a bit unfortunate. I pick the posters that I want to see from the program, but I also like to walk around the aisles and talk to authors of posters that just seem interesting. At ACL, this was problematic because of incredibly huge poster sessions which were incredibly crowded. It was also very loud.

The last aspect was much better at COLM (because of the venue with lower ceilings and carpet in the rooms). The sessions were also not that huge. They were still too crowded, and there was too little space around each poster, but: I actually perceived it differently than in Vienna: At COLM, I was happy that the posters were not thematically clustered, because I could just talk to the authors of 3-6 posters in each session that I found interesting. If they were clustered, I would have had very stressful and very boring sessions. Just one wish: put posters that are thematically similar closer to each other, such that one doesn't need to walk that far. But as the venue wasn't huge, that was also not such a big deal.

But generally, I still wish for:

- more space around each poster in poster sessions
- thematically similar posters should be close to each other

Altogether, this was not a small conference at all, but it felt much more manageable than the last ACL that I attended in Vienna. (EACL in Morocco and LREC in Mallorca were smaller and more manageable anyway)

Another interesting aspect I found was that there was no social event anywhere. Companies invited to socials, but there was no evening dinner or welcome reception. I really, really missed that. I didn't know as many people at this conference as I do know in the CL/NLP world, and I would have really liked to get to know more people who I didn't know yet, but that was more difficult than expected. In this regard, COLING is my favorite place, by the way, which typically offers trips in the middle of the conference during which you definitely start to talk to random people. I would wish COLM to adopt that concept.

Conference Program

I need to start with our own contributions in this blog post, because.

We had just one paper, and it was the first time (if I recall correctly) that we submitted to COLM. It's the work by Sean Papay (and me) [Papay and Klinger \(2026\)](#), in which we

propose a decoding algorithm that can consider lexical constraints; words that need to appear in the output. Current methods are either computationally efficient, or prefer constraint words that are more common in a language to appear first in the output. Our method is fair regarding this aspect and computationally efficient.

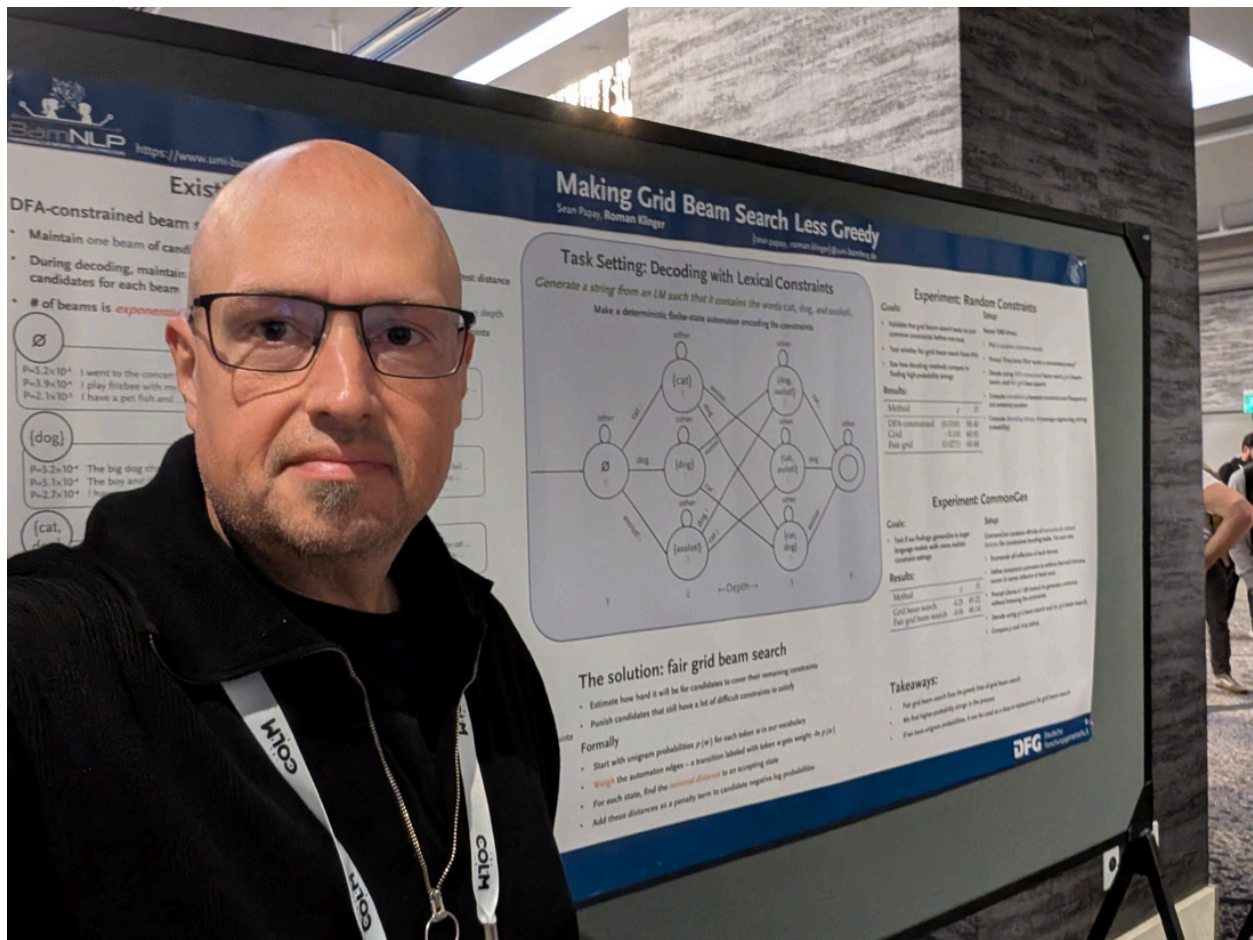


Figure 2: Me at the poster

Now, let's talk more about the program, which is, particularly for this conference which is different than others I attended so far, interesting. The papers are all available at [OpenReview](#). To get an overview, I downloaded all paper titles and ran a topic model (non-negative matrix factorization on one-hot vectors, labeled through an LM afterwards):

- LM agents, self evolution
- RL for LMs, verifiable rewards
- Decoding and generation efficiency
- Pretraining data
- Interpretability
- Safety and attacks
- Mixture-of-experts, compression, unlearning
- Multimodal and visual reasoning

This topic list also reflects my own impression of what dominated the conference. Nevertheless, I will not follow this categorization in my own paper overview, because that

topic list does not reflect too well my own interests. Instead, I will follow these categories (which do NOT reflect what has been popular at the conference, but purely my own research interests). Before we do that, though, let's talk about the invited talks.

Invited Talks

The conference featured 5 invited talks, and 1 panel. The topics were diverse, and all of them were interesting in some way. I want to mention three of the invited talks (I did not attend all of them):

Jennifer Pan: Neutrality Without Neutral Models



Figure 3: Jennifer Pan during her keynote

Jennifer Pan (Stanford University) started from the claim that political neutrality in AI is impossible, and asked what this means for the effect of AI on political persuasion. Her perspective comes from her research on information control in China: biased models can persuade, just like propaganda does in authoritarian regimes, but persuasion requires exposure. In competitive information environments, exposure is contested,

and research on political campaigns consistently finds small or even null effects of advertising and canvassing, despite enormous resources spent on them. Her conclusion was that it is exposure, not persuasiveness, that limits media effects. Neutrality might therefore be unattainable for any single model, but it can be approximated at the level of the ecosystem, through a diversity of models, viewpoints, and sources. What the Chinese government does, both in controlling digital information and now in generative AI, is to restrict exactly this diversity. The central political risk of AI is therefore not that a single model is biased, but that the whole ecosystem becomes politically homogeneous.

John Langford: Research at AI Speed



Figure 4: John Langford during his keynote

John Langford (Microsoft Research) talked about what he calls the industrialization of innovation, and how it changes the process and the outcomes of research and development. Given the title, I expected a talk about research with (or on) LLMs and AI. The talk was, however, essentially about how to speed up coding and software development, such that one can be more efficient with experimentation. It was quite industry focused. That was fine, and there were certainly useful thoughts in it – it was just not what I had expected.



Figure 5: Jason Eisner during his keynote

I was particularly looking forward to this talk. Jason Eisner (Johns Hopkins University) was a star that I really admired during my PhD, for his work on probabilistic graphical models. As this topic is much less prominent in many conferences these days, I was very curious to see how his work developed. And indeed, I could see in his talk that his research is still linked to these roots: it's all about factorizing probabilities.

His starting point was that everybody is building and evaluating LLM workflows today, but that we do not have an engineering discipline yet to find the most accurate and cost-effective system. There are many ways to break a task into subtasks, and many components to choose from – prompts, examples, rewards, LLMs, external tools, and humans in the loop, before the loop (annotators) or after the loop (evaluators). Do we need to search this space by hand? His proposal is to view it as active feature acquisition: every human annotation, LLM output, tool call (or, in medicine, every test result) is a random variable that we could pay to observe. As these variables are correlated, observing cheap ones helps to predict the expensive ones and to decide if they are worth observing at all. Gathering information therefore becomes a reinforcement learning problem. He described how to train an environment model that predicts distributions over any missing variables, similar to how BERT predicts missing tokens, and how to extend it into a policy that decides which variable to observe next. A nice

property is that observing a variable does not change the values of other variables, so the system can also learn from counterfactual trajectories.

My Favorite Contributions

Modelling of Psychological Concepts

(the summary sentences are generated by an LM)

- [Shi \(2026\)](#) investigates whether LLMs represent rule-based and consequence-based moral reasoning in distinct internal structures or merely react to surface wording, based on a factorial dataset built from established moral reasoning benchmarks.
- [Keluskar, Bhattacharjee, and Liu \(2026\)](#) manipulate the personality traits of agents in multi-agent LLM teams to study if personality composition affects task performance in coding, research collaboration, and competitive settings.
- [W. Li et al. \(2026\)](#) show that safety evaluations of persona-imbued LLMs need both prompting and activation steering, because the two methods reveal different, architecture-dependent vulnerabilities.
- [Merzouk et al. \(2026\)](#) propose to adversarially train on 66 harm-agnostic statements from the psychometric personality literature as a cheap alternative to training on harmful prompts to make models robust against jailbreaks.
- [Xiao et al. \(2026\)](#) introduce a privacy-themed theory-of-mind challenge in which a defender acts as a double agent who needs to make an attacker believe that sensitive information has been extracted successfully.
- [A. Lee et al. \(2026\)](#) introduce a multilingual, culturally adapted moral decision-making benchmark together with inference-time and training methods grounded in moral theory.
- [YS et al. \(2026\)](#) present a benchmark of multi-party scenarios that tests, based on the concept of theory of mind, whether LLM agents disclose private information appropriately when agents have different knowledge.
- [Getachew and Saparov \(2026\)](#) generate controllable synthetic stories to evaluate theory of mind and world modeling capabilities of LLMs with false-belief, goal, and preference attribution tasks, while avoiding contamination from pretraining data.
- [Y. Li et al. \(2026\)](#) evaluate the simulation of personalities with LLMs against what 1,000 public personalities actually said in 23,000 interview transcripts.
- [Sachdeva and Nuenen \(2026\)](#) let multiple LLMs debate the assignment of blame in 1,000 dilemmas from Reddit's "Am I the Asshole" community and show that the debate protocol influences the moral judgments and the consensus.

Emotion Analysis

- [Choi and Weber \(2026\)](#) use emotions, with their categorical and dimensional structure known from psychology, as a testbed to study the geometry of latent representations in LLMs.
- [Dokme, Reichman, and Heck \(2026\)](#) rewrite quantitative reasoning problems into emotionally framed variants (frustration, urgency, enthusiasm) with identical numerical content to test if the emotional framing alone degrades reasoning.

- [Ho et al. \(2026\)](#) induce emotions in LLMs with an imagination-based procedure and study with the Iowa Gambling Task if these emotions bias sequential decision making.

Biomedical and Clinical Text Mining

- [Chen et al. \(2026\)](#) study which aspects of a patient’s state can actually be recovered from transcripts of clinical encounters, using patient-reported outcome measures as an external anchor.
- [J. Wang et al. \(2026\)](#) provide a live, contamination-free platform for predicting the outcomes of clinical trials before they become public.
- [Hsu et al. \(2026\)](#) study active multi-turn diagnosis, in which models need to order tests and update their differential diagnosis, and analyze the failure modes of current LLMs in this setting.
- [Varadarajan et al. \(2026\)](#) evaluate whether LLMs adapt to cultural norms that users signal only implicitly, with health-related dialogues of personas from six cultures as a case study.
- [Byun et al. \(2026\)](#) study the confidence calibration of vision-language models in medical visual question answering, find persistent overconfidence, and propose a hallucination-aware mitigation.
- [Liu et al. \(2026\)](#) build a multi-turn adversarial benchmark in Japanese, based on guidelines of the Japan Medical Association, to test if LLMs stay aligned with medical ethics under jailbreak attempts.
- [Koretsky et al. \(2026\)](#) introduce a text-to-SQL benchmark over a biomedical knowledge base in which questions require implicit scientific reasoning.
- [Chan et al. \(2026\)](#) collect plain-language summaries of health research which are personalized by LLMs for lay readers and assess whether personalization helps comprehension and what it costs regarding safety.
- [Han et al. \(2026\)](#) present a benchmark with an interactive patient simulator in which models need to elicit concerns that patients do not disclose by themselves.
- [Zhou et al. \(2026\)](#) treat eligibility criteria of clinical trials as hard constraints in a retrieval system to match patients to trials with high recall.
- [Tang et al. \(2026\)](#) introduce a benchmark in which vision-language models need to report the differences between two temporally separated CT scans of the same patient.
- [J. Zhang et al. \(2026\)](#) evaluate whether vision-language models can reason about changes in a patient’s condition from medical images of different clinical visits.

Social Media Mining and Computational Social Science

- [Bouleimen, Pagan, and Hannak \(2026\)](#) test if LLM agents in social media simulations show the illusory truth effect (repeated exposure to a claim increases its perceived truth), which they would need to be valid proxies of human behavior.
- [Xing et al. \(2026\)](#) continually pretrain LLMs on engagement-heavy or low-quality Twitter/X posts and show that such “junk” data causes a lasting decline in reasoning, long-context understanding, and safety.

- [X. Zhang et al. \(2026\)](#) propose a multilingual agent-based simulation framework to model public opinion and media influence across languages and cultures.

Prompting and Instructions

- [Y. Wang et al. \(2026\)](#) show that prompts engineered for one model often degrade substantially when reused with another one and propose a training-free framework to transfer prompts across models.
- [Geng et al. \(2026\)](#) measure how pragmatic framing in instructions, such as “This is urgent” or “As your supervisor”, changes the behavior of LLMs although the task itself remains the same.

Bias, Safety, and Alignment

- [Hua, Lotfi, and Chen \(2026\)](#) show in a study of 50 quantized models that quantization flips up to 21% of individual responses between biased and unbiased, while aggregate bias scores remain unchanged.
- [B. W. Lee, Yueh-Han, and Korbak \(2026\)](#) do not attempt to prevent misbehavior of AI agents, but train them to produce a visible signal whenever they covertly misbehave.

Multimodal Models

- [Park et al. \(2026\)](#) structure the reasoning of vision-language models about social situations into cognitively motivated stages, from perception over situational understanding to norm-grounded judgment.
- [Feng et al. \(2026\)](#) show that scalar aesthetic scores poorly reflect the comparative preferences of experts and propose a benchmark to evaluate whether multimodal models can judge the beauty of images.

Other

- [Gao et al. \(2026\)](#) automatically organize large amounts of scientific literature into hierarchical structures that span multiple levels of abstraction, from broad domains to specific studies.
- [Elguendouze et al. \(2026\)](#) formulate argument component detection as a generative task, in which instruction-tuned LLMs jointly extract and classify claims and premises.
- [Bao and Carpuat \(2026\)](#) quantify reasoning errors of large reasoning models in machine translation across seven language pairs and study whether correcting these errors improves the translations.
- [Yi et al. \(2026\)](#) propose to post-train models for agentic tasks with reinforcement learning on existing supervised fine-tuning trajectories, which combines the compute efficiency of fine-tuning with the out-of-domain robustness of end-to-end reinforcement learning.

- [Shaib et al. \(2026\)](#) develop, based on interviews with experts, a taxonomy and interpretable dimensions to measure low-quality AI-generated text (“slop”).

Venue and Place

San Francisco is a great place for a conference. The city is compact enough that one can walk almost everywhere, there are endless options for food and coffee in walking distance to the venue, and the public transport works well enough to get to the airport and back without any trouble. What I missed a lot was some organized social event. I didn’t know too many people from the conference, and therefore, I spent a couple of evenings alone (fine for me, but not the conference experience I wished for).

The conference itself took place in a standard conference hotel. That is not the nicest place for a conference, to be honest: ballrooms without daylight, long corridors, carpet everywhere, and poster sessions squeezed into rooms. It works, it is practical, and the acoustics were (thanks to the carpet) much better than in Vienna. The city makes up for it, though.

One thing I noticed was that apparently a set of people were not there. In each poster session, some poster boards remained empty, and I also heard from others that authors could not make it. I do not know the reasons in each case, but I assume that travel to the US, visa issues, and costs played a role for some of them. That is unfortunate, both for the authors who could not present their work, and for the participants who wanted to talk to them. Still, altogether, San Francisco was a very good choice for COLM, and I would happily come back.

Bibliography

- Bao, Calvin, and Marine Carpuat. 2026. “Should We be Pedantic About Reasoning Errors in Machine Translation?” In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=PQLYIZzpT>.
- Bouleimen, Azza, Nicolò Pagan, and Ancsa Hannak. 2026. “Illusory Truth or Mere Exposure? Model-Dependent Repetition Effects in LLM-Based Social Media Simulations.” In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=S3vYJrl2Hn>.
- Byun, Ji Young, Young-Jin Park, Jean-Philippe Corbeil, and Asma Ben Abacha. 2026. “Overconfidence and Calibration in Medical VQA: Empirical Findings and Hallucination-Aware Mitigation.” In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=JH5MAXhj4v>.
- Chan, Joey, Yikun Han, Jingyuan Chen, Samuel Fang, Lauren D. Gryboski, Alexandra Lee, Sheel Tanna, et al. 2026. “ReLay: Personalized LLM-Generated Plain-Language Summaries for Better Understanding of Health Information, but at What Cost?” In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=O24T5DpwS2>.
- Chen, Lily, Ted Mau, Michael Gensheimer, Brian Anthony Nuyen, Nancy Jiang, and James Zou. 2026. “Conversation as Measurement in Clinical Encounters: Observable Phase Structure, Partially Observable Patient State.” In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=IZqvbZOX3I>.

- Choi, Benjamin J., and Melanie Weber. 2026. "Latent Structure of Affective Representations in Large Language Models." In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=inTQgPxkfw>.
- Chomsky, N. 1956. "Three Models for the Description of Language." *IEEE Transactions on Information Theory* 2 (3): 113–24. <https://doi.org/10.1109/TIT.1956.1056813>.
- Dokme, Atahan, Benjamin Reichman, and Larry Heck. 2026. "TEMPER: Testing Emotional Perturbation in Quantitative Reasoning." In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=g3uaspFhbA>.
- Elguendouze, Sofiane, Erwan Hain, Elena Cabrio, and Serena Villata. 2026. "Extracting Arguments, Not Just Classifying Them: Instruction-Tuned LLMs for Generative Component Detection." In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=pHotb9GjzO>.
- Feng, Yichen, Yuetai Li, Chunjiang Liu, Fengqing Jiang, Yue Huang, Yuanyuan Chen, Hang Hua, et al. 2026. "Visual Aesthetic Benchmark: Can Frontier Models Judge Beauty?" In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=WVtXQrM8bk>.
- Gao, Muhan, Jash Shah, Weiqi Wang, Kuan-Hao Huang, and Daniel Khashabi. 2026. "From Papers to Panoramas: Building Hierarchies of Scientific Literature at Scale." In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=GnSMhLhAgt>.
- Geng, Yilin, Omri Abend, Eduard Hovy, and Lea Frermann. 2026. "Measuring Pragmatic Influence in LLM Instructions." In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=DYIkaJDPIf>.
- Getachew, Nathaniel, and Abulhair Saparov. 2026. "Language Models Might Not Understand You: Evaluating Theory of Mind via Story Prompting." In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=DF3nbW0MCg>.
- Han, Yikun, Joey Chan, Jingyuan Chen, Mengting Ai, Simo Du, and Yue Guo. 2026. "MedConceal: A Benchmark for Clinical Hidden-Concern Reasoning Under Partial Observability." In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=rSTkVpdK5R>.
- Ho, Minh Khoi, Zihao Zhu, Runchuan Zhu, Levina Li, Zhiwen Fan, Zhangyang Wang, and Junyuan Hong. 2026. "Can Induced Emotion Bias LLM Behaviors in Sequential Decision Making?" In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=nDTN5W6hog>.
- Hsu, Hsin-Ling, Zizheng Wang, Donghua Zhang, Nai-Chia Chen, Jerry Wang, Jun-En Ding, Chia-Hsuan Hsu, et al. 2026. "MedAction: Towards Active Multi-turn Clinical Diagnostic LLMs." In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=2ZtsUgwVbx>.
- Hua, Stanley Bryan Zamora, Sanae Lotfi, and Irene Y. Chen. 2026. "Investigating Social Bias Changes in Quantized Large Language Models." In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=sr6xTPBZ5b>.
- Jurafsky, Daniel, and James H. Martin. 2026. "N-Gram Language Models." In *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3rd ed. <https://web.stanford.edu/~jurafsky/slp3/3.pdf>.
- Keluskar, Aryan, Amrita Bhattacharjee, and Huan Liu. 2026. "When Does Personality Composition Matter for Multi-Agent LLM Teams?" In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=MS8jsgZC62>.
- Koretsky, Mathew J., Maya Willey, Owen Bianchi, Chelsea X Alvarado, Tanay Nayak, Nicole Kuznetsov, Sungwon Kim, Michael Nalls, Daniel Khashabi, and Faraz Faghri.

2026. "BiomedSQL: Text-to-SQL for Scientific Reasoning on Biomedical Knowledge Bases." In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=SC7liryuqr>.
- Lee, Ayong, Ryan Sungmo Kwon, Yunxiang Zhang, Yuxuan Liu, Peter Railton, and Lu Wang. 2026. "MET: Theory-Grounded and Culture-Aware Multilingual Moral Reasoning." In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=pkazYtbTOg>.
- Lee, Bruce W., Chen Yueh-Han, and Tomek Korbak. 2026. "Training Agents to Self-Report Misbehavior." In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=rOZOqvZMGg>.
- Li, Wenkai, Fan Yang, Shaunak A. Mehta, and Koichi Onoue. 2026. "Persona Non Grata: Single-Method Safety Evaluation Is Incomplete for Persona-Imbued LLMs." In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=dyKvBEOToq>.
- Li, Yu, Pranav Narayanan Venkit, Yada Pruksachatkun, and Chien-Sheng Wu. 2026. "InterviewSim: A Scalable Framework for Interview-Grounded Personality Simulation." In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=7DYXGYDJPB>.
- Liu, Junyu, Zirui Li, Qian Niu, Zequn Zhang, Xun Yue, Wenlong Hou, Shujun Wang, Yusuke Iwasawa, Yutaka Matsuo, and Kan Hatakeyama-Sato. 2026. "JMedEthicBench: A Multi-Turn Adversarial Benchmark for Japanese Medical Ethics Alignment in LLMs." In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=ETOqcPMqpX>.
- Merzouk, Mohamed Amine, Nolan Smyth, Damiano Fornasiere, Linh Le, David Williams-King, and Adam Oberman. 2026. "Efficient Safety Alignment of Language Models via Latent Personality Traits." In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=TyejBOTa4N>.
- Papay, Sean, and Roman Klinger. 2026. "Making Grid Beam Search Less Greedy." In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=bapwJCTirq>.
- Park, Eunkyung, Wesley Hanwen Deng, Gunhee Kim, Motahhare Eslami, and Maarten Sap. 2026. "Cognitive Chain-of-Thought (CoCoT): Structured Multimodal Reasoning about Social Situations." In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=a0Z130ECIf>.
- Sachdeva, Pratik S., and Tom van Nuenen. 2026. "Interaction Protocol Shapes Moral Judgment in Multi-Agent Debate." In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=Th17bHGtz>.
- Shaib, Chantal, Tuhin Chakrabarty, Diego Garcia-Olano, and Byron C Wallace. 2026. "Measuring AI 'Slop' in Text." In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=UMipfwK6Ar>.
- Shi, Eugenie. 2026. "Rule vs. Consequence: Dissociable Internal Representations of Moral Reasoning in LLMs." In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=fvLg6Ek984>.
- Tang, Kegeng, Jingbo Wang, Shaogang Ren, and Zihao Wang. 2026. "CT- Δ Bench: A Benchmark for Longitudinal 3D Medical Imaging Difference Reporting with Vision-Language Models." In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=HsJP8pJNwq>.
- Varadarajan, Vasudha, Akhila Yerukola, Mona T. Diab, and Maarten Sap. 2026. "CCBench: Assessing LLM Cultural Competence via Implicitly Signaled Norms using Health Queries." In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=CCBench>.

- [//openreview.net/forum?id=fd4HmfxHZZ](https://openreview.net/forum?id=fd4HmfxHZZ).
- Wang, Jianyou, Youze Zheng, Longtian Bao, Hanyuan Zhang, Qirui Zheng, Yuhan Chen, Yang Zhang, et al. 2026. "CT Open: An Open-Access, Uncontaminated, Live Platform for the Open Challenge of Clinical Trial Outcome Prediction." In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=pbGCyTrXsl>.
- Wang, Yaxuan, Quan Liu, Zhenting Wang, Zichao Li, Wei Wei, Yang Liu, and Yujia Bao. 2026. "PromptBridge: Cross-Model Prompt Transfer for Large Language Models." In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=klFWV2uoNr>.
- Xiao, Hanqi, Vaidehi Patil, Zaid Khan, Hyunji Lee, Elias Stengel-Eskin, and Mohit Bansal. 2026. "Playing Along: Learning a Double-Agent Defender for Belief Steering via Theory of Mind." In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=zZU5EsxZ3S>.
- Xing, Shuo, Junyuan Hong, Yifan Wang, Runjin Chen, Zhenyu Zhang, Ananth Grama, Zhengzhong Tu, and Zhangyang Wang. 2026. "LLMs Can Get 'Brain Rot': A Pilot Study on Twitter/X." In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=FPDBbPDDvx>.
- Yi, Junkeun, Damon Mosk-Aoyama, Baihe Huang, Ritu Gala, Charles Wang, Sugam Devare, Khushi Bhardwaj, et al. 2026. "PivotRL: High Accuracy Agentic Post-Training at Low Compute Cost." In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=Ra73M7Xgbc>.
- YS, Yashwanth, Ruichen Wang, Shihua Zeng, Xuhui Zhou, Koichi Onoue, Vasudha Varadarajan, and Maarten Sap. 2026. "SOTOPIA-TOM: Evaluating Privacy and Information Management in Multi-Agent Interaction with Theory of Mind." In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=OjKywMsIOh>.
- Zhang, Junyi, Jia-Chen Gu, Wenbo Hu, Yu Zhou, Robinson Piramuthu, and Nanyun Peng. 2026. "TemMed-Bench: Evaluating Temporal Medical Image Reasoning in Vision-Language Models." In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=kyByX4OnSj>.
- Zhang, Xuan, Wenxuan Zhang, Anxu Wang, See-Kiong Ng, and Yang Deng. 2026. "Multilingual Agent-Based World Modeling for Social Science." In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=jteKCQPY5O>.
- Zhou, Zikai, Yufei Jin, Yilin Xu, Yu-Chiang Wang, Chieh-Ju Chao, and Monica Lam. 2026. "SatIR: Scalable High-Recall Constraint-Satisfaction-Based Information Retrieval for Clinical Trials Matching." In *Third Conference on Language Modeling*. <https://openreview.net/forum?id=bbOtBBwdrk>.

[[Download this post as PDF](#)]