ACL 2026 Industry Track: Overview

Yunyao Li[*]
Adobe
yunyaol@adobe.com

Georg Rehm[*]
DFKI
georg.rehm@dfki.de

Mei Tu[*]
Samsung
mei.tu@samsung.com

Abstract

For the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026), it was decided to organise a dedicated Industry Track once again. Similar to the main research track of the conference, the industry track attracted an unprecedented number of 532 paper submissions. In total, 419 reviewers and 28 area chairs participated in the evaluation of these papers. After a thorough, double-blind peer-review evaluation with at least three reviews for each submission followed by reviewer discussions and additional deliberations, 153 papers were selected for presentation at the ACL 2026 Industry Track. The submissions received can be grouped into six different clusters: 1. RAG systems & enterprise knowledge AI; 2. agentic systems and workflows; 3. model development, adaptation, optimization; 4. evaluation, benchmarking, LLM assessment; 5. safety, trustworthiness, responsible AI; 6. multimodal, speech, rich content processing.

1 Introduction

Language technologies and their applications are an integral and critical part of our daily lives. Many of these technologies have their roots in academic and industrial research laboratories where researchers invented a plethora of algorithms, benchmarked them against shared datasets and perfected their performance to provide plausible solutions to real-world applications. While a controlled laboratory setting is vital for a deeper scientific understanding of the problems underlying language technologies and the impact of algorithmic design choices on their performance, transitioning the technology to real-world applications raises a different set of challenging technical issues.

We acknowledge the challenges in adapting language technologies for building novel and robust real-world applications, as the journey from theoretical research to practical deployment can be difficult. Challenges can include technical aspects of system deployment and optimizing for efficiency, making informed design choices, or methodological considerations of incorporating human feedback and oversight. The Industry Track provides a forum to address these multifaceted issues. We were seeking submissions that not only delve into research, but also demonstrate the application of systems in real-world scenarios, irrespective of whether they involve proprietary data.

2 Call for Papers

We invited submissions that describe innovations and implementations in all areas of speech and natural language processing (NLP) technologies and systems that are relevant to real-word applications. The primary focus of the ACL 2026 Industry Track was on papers that advance the understanding of, and demonstrate the effective handling of, practical issues related to the deployment of language processing or generation technologies, including those of large language models, in non-trivial real-world systems, meaning: applications deployed for real-world use, i. e., outside controlled environments such as laboratories, classrooms or experimental crowd-sourced setups, also including applications that use NLP and/or speech technology, even if not state of the art in terms of research. There was no requirement that the system be made by a for-profit company, but the users of the system are most likely outside the NLP research community.

This track provided an opportunity to highlight key insights and new research challenges that arise from real-world implementations.

Relevant areas included system design, efficiency, maintainability, and scalability of real-world applications, with topics including, but not limited to (in alphabetical order):

[*]The three authors contributed equally to this overview article, to the preparation of the proceedings volume and to the organisation of the ACL 2026 Industry Track.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pages 1–4

- Benchmarks and methods for improving the latency and efficiency of systems
- Continuous maintenance and improvement of deployed systems
- Efficient methods for training and inference
- Enabling infrastructure for large-scale deployment
- Handling unexpected user behavior
- Human-in-the-Loop approaches to application development
- Implementation at speed, scale, and low-cost
- Negative results related to real-world applications
- System combination

Novel applications and use cases, with topics including, but not limited to (in alphabetical order):

- Best practices and lessons learned
- Case studies, from design to deployment
- Description of an application or system
- Design of application-relevant datasets
- Development of methods under system constraints (model or data size)
- Novel, previously unsolved NLP problems and novel NLP applications

Methods for deployed systems, with topics including, but not limited to (in alphabetical order):

- Ethics, bias, fairness, harmlessness, and trustworthiness in deployed systems
- Interpretability
- Interactive systems
- Offline and online system evaluation methodologies
- Online learning
- Robustness

Submissions had to clearly identify one of the following three areas into which they fall:

Deployed Must describe a system that solves a non-trivial real-world problem. The focus may include describing the problem related to actual use cases, its significance (against opportunity size, value proposition, and ideal end state), design/formulation of methods, tradeoff design decision for solutions, deployment challenges, and lessons learned.

Emerging Must describe the development of a system that solves a non-trivial real-world problem (it need not be deployed or even close, but there needs to be evidence that this development is intended for real-world deployment). Papers that describe enabling infrastructure for large-scale deployment of NLP techniques also fall in this category.

Discovery Must include results obtained from NLP applications in real-world scenarios that result in actionable insights. These discoveries should reveal promising directions in their application areas, leading to further system or societal enhancements. For example, an actionable discovery from an analysis of call center transcripts may reveal that certain language choices negatively impact customer experience, leading to better training of service representatives and improved customer experience.

3 Submissions and Results

The call for ACL 2026 Industry Track papers attracted an unprecedented number of 532 submissions. A total of 419 reviewers and 28 area chairs participated in the evaluation of these submissions. After a thorough double-blind peer-review evaluation with at least three reviews for each submission, we eventually selected a total of 153 articles for presentation within the ACL 2026 Industry Track, with 61 oral and 92 poster presentations (acceptance rate: 34.16%).

4 Research Trends

As expected, almost all submissions revolve around LLMs, indicating the prevalence of their adoption in real-world applications. More specifically, we can cluster the submissions received in 2026 into the following six groups of broader topics.

RAG Systems & Enterprise Knowledge AI This cluster centers on the integration of language models with external knowledge sources in order to produce reliable, verifiable outputs in real-world settings. The papers in this group typically focus on retrieval-augmented generation (RAG), enterprise search, and question answering systems that operate over large, heterogeneous document collections. A recurring concern is how to ground model outputs in authoritative sources such as legal texts, clinical data, or internal enterprise knowledge bases, thereby improving factual accuracy and traceability.

Taken together, this line of work reflects a shift from standalone generative models toward systems that are tightly coupled with structured and unstructured knowledge, enabling more trustworthy and domain-aware applications.

Agentic Systems and Workflows The papers in this cluster treat language models not merely as text generators but as active components within larger computational workflows. The focus is on agentic architectures in which LLMs can plan, decompose tasks, interact with external tools and APIs, and coordinate with other agents. Typical contributions explore multi-agent systems, tool use, memory mechanisms, and execution monitoring in complex environments. These systems are often designed for practical deployment, where they automate multi-step processes or assist users in accomplishing structured tasks. The overarching theme is the transformation of LLMs into decision-making entities that can operate within and orchestrate real-world processes rather than simply producing isolated outputs.

Model Development, Adaptation, Optimization This cluster brings together work on adapting LLMs to the practical constraints of industrial deployment. The emphasis lies on methods that improve efficiency, scalability, and domain relevance, including fine-tuning, instruction tuning, distillation, pruning, and quantization. Many papers also address inference-time optimization, model routing, and cost–latency trade-offs in production environments. Rather than proposing entirely new model architectures, the contributions in this area focus on making existing models usable at scale, under resource constraints, and in specialized domains. As such, this cluster reflects a maturing phase of the field, where the central challenge is no longer building models per se, but making them performant and economically viable in real-world systems.

Evaluation, Benchmarking, LLM Assessment The papers in this group concern with how to measure the performance and usefulness of LLMs in realistic scenarios. They explore new evaluation frameworks, benchmark datasets, and metrics that go beyond traditional academic settings, often incorporating human judgment or task-specific criteria. A notable trend is the use of LLMs themselves as evaluators, alongside more structured approaches to assessing reasoning, generation quality, and code performance. The underlying motivation is to develop evaluation methodologies that better reflect deployment conditions and user expectations. This cluster highlights the growing recognition that robust, meaningful evaluation is a prerequisite for reliable and trustworthy AI systems.

Safety, Trustworthiness, Responsible AI This cluster addresses the challenges of ensuring that language models behave in safe, predictable, and policy-compliant ways when deployed in real-world contexts. The work spans a range of topics, including hallucination detection and mitigation, content moderation, privacy preservation, bias reduction, and robustness against adversarial inputs. Many contributions also consider regulatory and governance aspects, reflecting the increasing importance of compliance in industrial applications. The common thread is the need to align model behavior with human expectations, legal requirements, and ethical standards. This area underscores that technical performance alone is insufficient; systems must also be trustworthy and controllable.

Multimodal, Speech, Rich Content Processing The final cluster extends beyond text-only processing to encompass multimodal and content-rich data. The papers in this area explore the integration of language models with speech, audio, images, video, and complex documents. This includes work on speech recognition and understanding, document AI (such as OCR and structured information extraction), and multimodal retrieval or generation systems. The goal is to enable AI systems that can operate on the diverse data types encountered in real-world applications, from scanned forms to multimedia content. This cluster reflects the broadening scope of language technologies as they evolve into more general-purpose interfaces for

interacting with heterogeneous information sources.

The submissions received for the ACL 2026 Industry Track demonstrate that language technology is moving toward grounded, agentic, efficient, evaluated, safe, and multimodal systems that are deeply embedded in real operational environments. We expect these trends to continue and rapidly evolve in the near future.

5 Invited Keynote

Roberto Navigli (Sapienza University of Rome and Babelscape, Rome, Italy)

6 Programme Co-Chairs

- Yunyao Li, Adobe, USA
- Georg Rehm, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH and Humboldt-Universität zu Berlin, Germany
- Mei Tu, Samsung Research China, China

Acknowledgments

The ACL 2026 Industry Track Programme Co-Chairs would like to thank the authors of all submissions as well as the reviewers and area chairs for their hard and dedicated work under very tight deadlines. We would also like to thank the General Chair and ACL 2026 committees with which we interacted between the summer of 2025, when this endeavour started, and the summer of 2026, when we finally have been able to have the Industry Track at the ACL 2026 conference in San Diego, California, USA. Finally, we would also like to thank our keynote speaker, the ACL team, especially Jennifer Rachford, as well as the Underline team, especially Damira Mrsic.

Georg Rehm was supported through the project NFDI for Data Science and Artificial Intelligence (NFDI4DS) as part of the non-profit association National Research Data Infrastructure (NFDI e. V.). The NFDI is funded by the Federal Republic of Germany and its states.