

EMNLP 2026 Industry Track: Overview

Pepa Atanasova¹ Gholamreza Haffari² Minghan Wang²

¹University of Copenhagen, Denmark ²Monash University, Australia

pepa@di.ku.dk

{gholamreza.haffari, minghan.wang}@monash.edu

Abstract

For the 2026 Conference on Empirical Methods in Natural Language Processing (EMNLP 2026), a dedicated Industry Track was organised to provide a forum for work on real-world applications of speech and language technologies. The track received 931 submissions, which were evaluated by 733 reviewers and 158 area chairs. Following the review process, 265 papers were selected for presentation, including 23 oral presentations and 242 posters. Across the accepted programme, prominent themes include evaluation and benchmarking, agents and orchestration, efficient serving and cost control, retrieval and knowledge grounding, and safety and reliability. This overview summarises the scope of the track and its submission and review process, highlights these research trends, and describes the Industry Track programme, including its oral sessions and panel discussion.

1 Introduction

Language technologies are now widely used in real-world systems across many domains. While many advances originate in controlled research environments, deploying these technologies in practice raises additional challenges, including latency and cost constraints, changing user behaviour, maintenance over time, evaluation without clean test sets, integration with human workflows, and trade-offs between quality and practicality.

The EMNLP 2026 Industry Track was organised to provide a forum for work addressing these issues. We invited submissions on the design, development, deployment, and analysis of speech and language systems in real-world settings. Submissions were not restricted to industry authors or proprietary data; academic work on genuinely deployed systems, negative results, lessons learned, and vision papers grounded in deployment experience were also welcome.

2 Call for Papers

We invited submissions describing innovations and implementations in speech and language technologies for real-world applications. The primary focus was on work that advances the understanding of practical challenges arising in non-trivial real-world settings and demonstrates how such challenges can be handled in deployed systems or systems intended for deployment.

Relevant topics were organised into three broad areas:

System design, efficiency, maintainability, and scalability

- 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 behaviour
- Human-in-the-loop approaches to application development
- Implementation at speed, scale, and low cost
- Negative results related to real-world applications
- System combination and orchestration

Novel applications and use cases

- 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, such as model or data size
- Novel, previously unsolved NLP problems and novel NLP applications

Methods for deployed systems

- Ethics, fairness, safety, and trustworthiness
- Evaluation of deployed systems

- Interpretability and transparency
- Interactive systems
- Online learning and adaptation
- Robustness and reliability

Opinion and vision papers related to real-world applications were also welcome. Submissions were reviewed double-blind and assessed based on novelty, technical quality, potential impact, and clarity, with particular emphasis on real-world implementations, real-world data, and appropriate evaluation methodology. The Industry Track used a dedicated OpenReview submission and review process and did not use ARR.

Original submissions were limited to six pages of main content. Accepted papers were allowed one additional page of main content in the final version. The final version was required to remove anonymisation, and all accepted papers are to be presented at the conference either orally or as posters.

3 Submissions and Results

The call for EMNLP 2026 Industry Track papers attracted 931 submissions. A total of 733 reviewers and 158 area chairs participated in the evaluation of these submissions. Following the review process, 265 papers were selected for presentation, with 23 oral presentations and 242 poster presentations (acceptance rate: 28.5%).

4 Research Trends

While the industry tracks so far have collected a large amount of papers about *building* a language-model capability, the 2026 stream of work has overwhelmingly focused on *operating* with language models: deciding whether a system is good enough to ship, what it costs to run, how it fails when conditions change, and who is accountable when it does. The most common single keyword remains “large language models” (34 papers), but the next most frequent: “LLM agents” (20), “LLM-as-a-judge” (15), “benchmark” (14) and “evaluation” (11), name the concerns around deploying language models rather than a building further capabilities.

Clustering the accepted papers by the topics named in the title and keywords gives the distribution in Table 1¹. Papers can be located in more

than one cluster where they span multiple topics, which is common for most papers: each paper is in 2.3 clusters on average, and only 7 of 265 papers fall outside this distribution.

The most frequent co-occurrence of topics is *agents* with *evaluation* (37 papers), followed by *agents* with *efficiency* (21) and *agents* with *retrieval* (21). *Evaluation* naturally pairs with nearly every topic: *safety* (19), *retrieval* (18), *multimodal* (18), *vertical domains* (17).

Evaluation becomes a central topic. A third of the accepted papers are primarily about measurement, and the topic has the highest acceptance rate (43.1%). The papers are not proposing general-purpose leaderboards. They ask narrower, operational questions, e.g., whether an automatic judge can be trusted on a particular decision, when to escalate to a human, how an evaluation built offline behaves once traffic changes. The recurrence of “LLM-as-a-judge” (15 papers) suggests a field that is adopting model-based evaluation and is now auditing the auditor.

Agents are increasingly used in operation. Ninety-four accepted papers concern agents, tools or orchestration, with a shift in the focus. Rather than showing that an agent can complete a task, they address what happens on the thousandth run, e.g., skill and memory persistence, tool selection under a large and changing catalogue, protocol standardisation, architectural enforcement of permissions, and benchmarks that measure delivery.

the head of the distribution—“large language models” (34 papers), “LLM agents” (20), “LLM-as-a-judge” (15), “benchmark” (14), “evaluation” (11), “retrieval-augmented generation” (10), together with the long tail, and defined thirteen clusters that the recurring vocabulary suggested. Second, each cluster was expressed as a set of surface patterns over the concatenated *title and keywords* of each paper: for example, the efficiency cluster matches *quantiz, distill, lora, latency, throughput, serving, cost-aware, routing, cach, small language model and compress*. Third, assignment is multi-label. A paper matching several clusters is counted in each, which is why the column of accepted counts sums to more than 265. Acceptance rates are computed by applying the same patterns to the rejected submissions and dividing. This brings two limitations. The patterns are lexical, so a paper that describes cost control without using any of the efficiency vocabulary is missed; and cluster membership reflects the words authors chose, not an independent reading of the contribution. The seven unmatched papers cover plan repair, AI-text detection bias against non-native writers, RTL coding for hardware design, a document representation for LLM-assisted scientific writing, long-context aspect-based summarisation, pharmacological parameter extraction, and inventory-grounded search optimisation.

¹**How the clusters were built.** The authors of the accepted papers supplied 1,404 free-text keywords drawn from a vocabulary of 1,045 distinct strings, so almost every keyword is used once or twice and there is no taxonomy to group them. We therefore proceeded as follows. First, we ranked the keyword vocabulary by frequency and analysed

Cluster	Accepted	Share	Acc. rate
Evaluation, benchmarking and LLM judges	100	37.7%	43.1%
Agents, tools and orchestration	94	35.5%	36.0%
Efficiency, serving and cost control	68	25.7%	41.2%
Retrieval, RAG and knowledge grounding	65	24.5%	38.0%
Safety, guardrails and trust	46	17.4%	38.7%
Multimodal, speech and documents	45	17.0%	37.5%
Conversational and customer-facing systems	38	14.3%	46.9%
Post-training, RL and adaptation	36	13.6%	40.4%
Vertical domains: finance, legal, health	36	13.6%	36.7%
E-commerce, search and recommendation	30	11.3%	38.0%
Data curation and synthesis	19	7.2%	47.5%
Code, structured output and APIs	18	6.8%	32.1%
Multilinguality and localisation	18	6.8%	34.0%

Table 1: Thematic clusters among the 265 accepted papers. Papers may appear in several clusters. “Acc. rate” is the share of submissions in that cluster that were accepted.

Cost becomes a central research variable.

Sixty-eight papers treat efficiency, serving or cost control as a primary topic, including e.g., quantisation, distillation, LoRA, routing, caching, small language models, unit economics. Notably, cost is rarely reported as an add-on to an performance table, but appears in the problem formulation as a constraint the method is designed to satisfy, e.g., cost-aware routing, certified cost-optimal triage, cost-efficient guardrails.

Reliability work is focusing on the architecture.

The safety, guardrails and trust cluster (46 papers) contains comparatively little prompt-level mitigation. The accepted work tends to place controls in the system rather than the model, e.g., proxies that enforce tool access, deferral routers that decide when not to answer, abstention and calibration mechanisms, unlearning with controllable scope, and measurement methods for failures (refusal aversion, overthinking, sycophancy). All these appear under real user traffic. This is the clearest signal that deployment experience, rather than benchmark performance, now sets the research agenda.

5 The Programme

The Industry Track programme combines six Industry sessions, a poster session, two invited talks, and a closing panel.

Industry 1: Opening, Retrieval and RAG (25 October) focuses on retrieval and grounding for deployed systems, including robust conversational search under distribution shift, vendor-agnostic grounding for LLM agents, continual correction in agentic RAG, structured-semantic conversation analytics, and tree-relational enhancement for oncology LLMs.

Industry 2 (25 October) opens with an invited talk by Mohit Bansal, followed by two oral presentations. The session includes *Deploying LLM Agents for Auditable Complaint Investigation in Banking*, which received the Industry Track award, and *WARP: Wasserstein-Aligned RAG for Population Opinions*.

Industry 3 (26 October) features two oral presentations on evaluation and benchmarking—margin-calibrated routing between LLM-as-a-judge and human experts, and multi-step tool calling over Korean open public APIs—followed by an invited talk by Verena Rieser. The Industry poster session follows later that day.

Industry 4: Trust, Safety and Control (26 October) brings together six oral presentations on architectural enforcement of tool access, causal measurement of refusal behaviour in deployed LLMs, overthinking in content moderation, structural interpretability of agentic reasoning traces, claim verification in investment-agent debates, and selective unlearning.

Industry 5: Agents in Production (27 October) focuses on operational agent systems, covering strategy evolution across multi-session sales interactions, capability discovery beyond in-context routing, governed multi-agent customer service, agentic response orchestration at scale, self-evolving tool-making agents, and state-dependent risk measurement for enterprise agents.

Industry 6 (27 October) closes the track with two oral presentations, the panel discussion, and closing remarks. *AutoGraphQL: A Bring-Your-Own-Schema Framework for Adapting Open-Weight LLMs to GraphQL APIs* and *ForensicBench: Evaluating Agentic LLMs on Journal-Entry Fraud*

Detection received the Industry Track runner-up awards.

The closing panel, *AI Agents in the Real World: Capability, Trust, Values, and Deployment*, brings together the programme's recurring questions around capability, trust, values, and the practical deployment of agentic systems.

6 Programme Co-Chairs

- Pepa Atanasova, University of Copenhagen, Denmark
- Gholamreza Haffari, Monash University, Australia
- Minghan Wang, Monash University, Australia

Acknowledgments

The EMNLP 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. We would also like to thank the EMNLP 2026 General Chairs and organising committees for their support throughout the preparation of the Industry Track.