

The LM Challenge 2026

Sherzod Hakimov[⊗], Dimitrije Ristic[⊗], Philipp Sadler[⊗], Davide Mazzaccara[‡],
Sabrina McCallum[◊], Filippo Momentè[‡], Maria Manina[⊗], Karl Osswald[⊗], Dieu-Thu Le[◊],
Oliver Lemon[◊], Alexander Koller[§], Raffaella Bernardi^{*}, Raquel Fernández[△],
Mario Giulianelli[×], Alessandro Suglia[◊], David Schlangen^{⊗, ⊕}

[⊗]University of Potsdam, [‡]University of Trento, [◊]University of Edinburgh,

[◊]Amazon AGI, Germany, [◊]Heriot-Watt University, [§]Saarland University,

^{*}Free University of Bozen-Bolzano, [△]University of Amsterdam, [×]UCL, [⊕]DFKI

lm-playschool@googlegroups.com

Abstract

Children acquire language from far less input than language models, and social interaction is widely thought to be part of the reason. Dialogue games cannot fully replicate human interaction, but they offer a controllable environment for simulating situated language use and for studying how models learn from and through interaction. The LM Playschool Challenge builds on this idea, shifting language model training away from observational data towards multi-turn language use. Participants were free to post-train a language model of any size, using any data or method, with one objective in mind — to improve its performance on an unseen set of dialogue games. Our findings suggest that post-training on dialogue games effectively teaches game rules and structure and can support generalisation, especially for small models — the best system achieved gains of roughly seven and three times its baseline score on held-out instances of seen games and on unseen games, respectively. However, gains on unseen games were concentrated in a few games, and larger models improved mainly by achieving better completion rates. General ability was largely retained across submitted systems, owing to the fact that fine-tuning on dialogue games produces shorter responses — and in turn, trigger fewer formatting errors on static benchmarks. Nearly all submissions relied on supervised fine-tuning, often followed by preference optimisation, and none trained on the dense turn-level feedback offered by Playpen, leaving this an open direction for future iterations of the challenge.

1 Introduction

The **LM Playschool Workshop** is concerned with studying and developing language agents that learn or improve through social interaction in conversational, collaborative, goal-oriented, multi-turn environments. The **LM Playschool Challenge** gave

this broad research programme a concrete empirical realisation by asking whether post-training large language models through *dialogue games*—i.e., goal-oriented tasks governed by multi-turn language interaction—can improve their interactive capabilities. This question is relevant to the development of more useful language technologies because real-world interactions *are* conversational, collaborative, goal-oriented, and multi-turn. It is also relevant to the development of models that learn and use language in more human-like ways, both with respect to the formal and functional skills (Mahowald et al., 2024) they acquire and to the efficiency of that learning.

Dialogue games provide a useful setting for evaluating language agents and their ability to use language in interaction. In collaborative dialogue games, a language agent works with a partner towards a joint goal (Schlangen, 2023). Success therefore depends on its ability to understand the task, coordinate with its partner through language, select appropriate actions, and adapt to the changing game state across multiple turns. Previous work has shown that LLMs can be prompted to play a wide range of interactive language-mediated games (Chalamalasetti et al., 2023; Li et al., 2023; Qiao et al., 2023; Xu et al., 2023). These games test verbal interaction skills that are not adequately covered by single-turn benchmarks. They can discriminate between models, correlate with cognitive tests targeting core executive functions (Momentè et al., 2025), and reveal that current LLMs still fall short of human performance (Beyer et al., 2024).

Dialogue games are also promising environments for training language agents because they provide an intrinsic learning signal derived directly from success or failure in the game (Wang et al., 2024a; Cheng et al., 2024; Gul and Artzi, 2024; Horst et al., 2025). First, this signal evaluates utter-

ances by how effectively they advance a shared goal rather than by their local appropriateness alone. It is also an objective signal, in the sense that it can be computed programmatically from the game state without relying on a learned judgment model, and it is grounded in interaction because task success depends on successful communication across turns. These properties support several approaches to post-training, including offline learning from recorded interactions and online learning through interaction with the game environment.

The study of how language agents learn through interaction is also informed by work on human language acquisition. Human children learn from far less linguistic input than current neural language models require (Hart and Risley, 1995; Cristia et al., 2019; Linzen, 2020), and social interaction is widely hypothesised to contribute to this data efficiency (Clark, 2016; Bloom, 2000; Bisk et al., 2020; Bender and Koller, 2020). Dialogue games do not reproduce the conditions of human language acquisition, but they provide a controlled setting in which to investigate whether learning through interaction can make model training more effective and data-efficient.

The LM Playschool Challenge brought together these three lines of work: language-based games as a means of evaluating large language models and their ability to use language in interaction; game-based and interactive environments for training large language models; and connections between human language acquisition and machine language learning, with particular attention to the role of social interaction in data-efficient learning. The challenge asked participants to improve language agents through model post-training. Participants were free to choose their base models, data, and training methods, but were particularly encouraged to use interaction data. The public Playpen environment (Horst et al., 2025) provided example games, learning scripts, and a development evaluation setup. Submitted systems were then evaluated on unseen dialogue games drawn from a private version of the clembench framework (Chalamalasetti et al., 2023) as well as on static evaluation benchmarks, with improvement measured relative to each system’s base model. The challenge received eleven systems from eight teams, all of which were built on Qwen3.5. Most submissions relied on offline supervised fine-tuning on Playpen interactions, often followed by DPO (Rafailov et al., 2023), while two submissions ex-

plored inference-time methods and only one used online reinforcement learning.

Across the submitted systems, post-training on dialogue games teaches models the rules and structure of the games more effectively than optimal strategies. The most significant gains were observed in the smallest (2B) models, where the best system achieved a 7-fold improvement on unseen instances of training games and a 3-fold on entirely unseen games relative to the baseline model, through gains in both game completion and game quality. These results demonstrate that post-training can support generalisation, especially for smaller models, although gains on unseen games were smaller than those in-domain and were concentrated on a small subset of games. For larger models, improvements were mainly in-domain and driven by higher completion rates on a wide range of games. On static benchmarks, despite a slight overall decline, post-training produced more concise outputs and fewer formatting errors.

These findings are based on the approaches explored by the submitted systems and therefore leave several promising directions open. In particular, no submission trained directly on the dense feedback that the Playpen provides within an episode, leaving learning from feedback during interaction as an interesting direction for future iterations of the challenge.

2 The Playpen Environment

Playpen is an interactive environment for learning from and in dialogue games. It supports the use of dialogue games both as evaluation devices and as environments for model training by structuring interaction around the standard observe-act loop and providing access to turn-level and episode-level rewards. It can be considered as a learning-oriented wrapper around the clembench package, which in itself focussed on evaluation and not learning. For the challenge, we re-wrote the environment used in (Horst et al., 2025) with a focus on ease-of-use and extensibility.¹

A dialogue game in clembench (and hence, in Playpen) involves one or more player agents whose interactions with the environment and with one another are mediated by a programmatic GameMaster. At each turn, an agent receives an observation containing the dialogue history together

¹The package is being developed further; the current state can be found at <https://github.com/lm-playpen/playpen>.

with any game-specific state that is available to its role, and produces an action. Depending on the game, the action may take the form of a move command, a clue, a guess, or a message to another player. The GameMaster processes the action, checks that it satisfies the rules of the game, updates the game state, and returns the next observation together with any associated reward. This process continues until the game reaches a terminal state.

Beyond standard game-specific reward signals, Playpen can also incorporate additional sources of feedback, such as responses from judgment models or verifiers, and provide agents with access to tools. Integration with the PettingZoo (Terry et al., 2021) and Gymnasium (Towers et al., 2025) provides interfaces for multi-agent and single-agent learning, respectively, while integration with OpenEnv sandbox (OpenEnv Technical Committee, 2025) supports isolated execution of game environments. Playpen is designed around the clembench suite of games (Chalamalasetti et al., 2023; Schlangen et al., 2026). Users can customise existing games, define task-specific scores and reward functions, generate new game instances, or design new games altogether.

Specifically for the challenge, we provided a simple command-line interface through which a candidate model could be validated on a specific set of game-instances held back from the training set, as well on some static benchmarks; as described in more detail in the next section. We also packaged up transcripts from previous runs of the benchmark (as documented at <https://clembench.github.io>) into an easy to process format, for experiments with learning from demonstrations.

3 Evaluation

The LM Playschool Challenge asked participants to post-train a baseline model and submit the resulting model checkpoint. (Note: What we call *baseline model* here are publicly released, instruction-tuned and post-trained models that form the basis for our *continued* post-training; not to be confused with base models that are only pre-trained.) We evaluate these checkpoints following Horst et al. (2025), comparing each post-trained model with the corresponding base model across three main dimensions. **In-domain performance** measures improvement on unseen game instances of the clembench games that were available during training. **Out-of-domain performance** is evaluated on previously

unpublished clembench games, providing a measure of how well improvements transfer to tasks that were not available during post-training. Finally, we evaluate whether models **retain their general capabilities** after post-training using a set of static benchmarks (e.g., MMLU-Pro, Wang et al., 2024b).

In addition to this main evaluation, we conduct separate validation runs on the game instances included in the official Playpen validation set and on the static benchmark tasks released to participants as part of the challenge package. All evaluation runs were carried out after the submission deadline using a fixed evaluation pipeline and the same configuration across submissions.²

3.1 In-Domain Evaluation

In-domain evaluation measures whether post-training improved a model’s gameplay on the distribution of games it was trained on, but using unseen game instances. The in-domain set consists of 16 games and game variants featured in the Playpen training set and available as part of clembench. The instances were produced following the standard recipe for the benchmark. The instances, together with the material described in the next section, are now available at <https://github.com/lm-playpen/lm-playschool-2026-closed-set>.

3.2 Static Benchmark Evaluation

Static benchmark evaluation assesses whether post-training improves or deteriorates general capabilities. We evaluate each model on 100 instances from each of five static benchmarks: BigBenchHard (BBH; Suzgun et al., 2023; Srivastava et al., 2023), CLadder (Jin et al., 2024), EQ-Bench (Paech, 2023), IFEval (Zhou et al., 2023), and MMLU-Pro (Wang et al., 2024b).

3.3 Out-of-Domain Evaluation

Out-of-domain evaluation measures whether improvements acquired through post-training transfer to games that are beyond the training distribution, or whether the model has overspecialised to the games available during training. For this purpose, we introduce a set of 13 new, previously unpublished games and game variants; hence, participants didn’t have access to them when they

²The full runs, including all transcripts of gameplay, are available at <https://github.com/lm-playpen/lm-playschool-2026-final-results>.

submitted their final checkpoints. Five are new games developed for the challenge (CHRONICLE, CLOCKWORK COURIER, CRYPTOLECT, TOWER OF HANOI, and GET TO THE POINT), two are new variants of games already available during training (CLEAN UP — SEMANTIC TEXT and WORDLE CRAZY), and four are adaptations of games from TextArena (Guertler et al., 2025) (SOKOBAN, BLACKJACK, FROZEN LAKE, and MASTERMIND).

CHRONICLE is a two-player historical deduction game. The game features a Narrator who describes a secret historical event (such as a battle, scientific discovery, or a political milestone) to a Detective over 10 turns, one sentence at a time, without naming the event or using forbidden event-specific nouns. The game evaluates a model’s world knowledge as well as the ability to describe and recognise an event from partial descriptions.

CLOCKWORK COURIER is a single-player planning and navigation game played on a dynamic map that changes according to a pre-defined timetable. A courier must collect every parcel on the map and deliver it to its designated destination before time runs out. Map navigation is further complicated by gates that open and close on a fixed timer and guards that patrol predefined routes. The full map and timetable are available to the player from the beginning, making the task one of planning under known, deterministic dynamics. The game assesses the ability to plan and timely execute actions over a dynamically changing environment.

CRYPTOLECT is a single-player grammar induction game in which the model assumes the role of a field linguist learning an unknown constructed language through interaction with a programmatic informant. The final task is to translate a set of previously unseen sentences into the new language. The game evaluates the ability to acquire information through interaction, infer latent rules from evidence, and generalise these rules to new examples.

GET TO THE POINT is a two-player word-association game where one player, the Helper, attempts to guide the other, the Seeker, from a fixed starting word towards a hidden target word through indirect clues. The game resembles Taboo, which is included in the in-domain set, but differs in its communication constraints. Rather than avoiding a list of forbidden words, the Helper may use the starting clue word and up to three additional words on each turn. The game therefore evaluates semantic reasoning as well as, given its similarity to

Taboo, a model’s ability to transfer skills acquired on a related training game to a task with different interaction rules.

TOWER OF HANOI is a text-based implementation of the classic puzzle, featuring three pegs and disks of varying sizes. The model must move all disks to a target peg, moving only one disk per turn and never placing a larger disk on top of a smaller one. The game evaluates state tracking and multi-step planning.

We additionally include two previously unseen variants of in-domain games. In CLEAN UP – SEMANTIC TEXT, two players are given a 7×7 ASCII grid containing letter objects at different positions. The goal is to rearrange the objects on their private grids so that identical objects occupy matching coordinates. Unlike in the original game, explicit coordinates and row or column instructions are forbidden, and players are required to describe locations relative to objects in the surrounding environment. In WORDLE CRAZY, the core game mechanics remain the same as in WORDLE, but the mapping between guesses and colour-coded feedback is altered. Models must therefore infer and adapt to the modified feedback convention rather than relying on the standard WORDLE rules.

Finally, SOKOBAN, BLACKJACK, FROZEN LAKE, and MASTERMIND are well-known games integrated from TextArena. They test several forms of general-purpose planning and reasoning: spatial planning, decision-making under uncertainty, navigation under stochastic dynamics, and logical deduction from constrained feedback, respectively.

3.4 Evaluation Pipeline

For evaluation, each author submitted a merged model checkpoint via HuggingFace, along with information about the baseline model it was fine-tuned from. Authors also had to submit an entry for the clemcore model registry, specifying the configuration to use during evaluation.³

In total, six teams submitted fine-tuned checkpoints, while two teams took part with inference-time approaches guided by process reward models.⁴ All fine-tuned submissions were evaluated with temperature set to 0 and a max token limit of 5000. For the two special submissions, we used the default token limits and temperatures provided in

³<https://github.com/clp-research/clemcore>

⁴The two inference-time approaches decided not to submit a technical report.

their code, and set the search budget to best-of-4 sampled responses due to resource constraints.

Evaluation was run on a cluster of 4 x A100 80GB GPUs provided by the University of Potsdam. We evaluated the submitted checkpoints, along with 5 baseline models from the Qwen 3.5 (Qwen Team, 2026) and Llama 3.1 (Grattafiori et al., 2024) families, consuming 415 GPU hours.

3.5 Metrics and Scoring

The main score featured in Playpen (following clembench) is the **Clemscore**. For each game, it combines *% Played*, the percentage of instances completed, with a game-specific *Quality Score*:

$$\text{Clemscore} = \text{Quality Score} \times \frac{\% \text{ Played}}{100}. \quad (1)$$

The resulting scores are macro-averaged across games. Performance on the five static benchmarks, instead, is averaged into the **Statscore**. We refer to performance on the clembench games and static benchmarks from the official Playpen validation subset as Validation Clemscore and Validation Statscore, respectively. Following Horst et al. (2025), we report the Clemscore separately for in-domain (ID) and out-of-domain (OOD) games and calculate their changes relative to the corresponding baseline models. We use these scores and their relative changes to derive three further metrics.

To combine our two objectives of improving gameplay while not regressing on general benchmarks, the **adjusted delta** (Δ_{adj}) corrects the change in Clemscore for any decline in static benchmark performance:

$$\Delta_{\text{adj}} = \Delta + \min(\Delta_{\text{Statscore}}, 0). \quad (2)$$

Here, Δ is the change in the ID or OOD game score relative to the baseline, and $\Delta_{\text{Statscore}}$ is the change in Statscore. A decline in Statscore reduces the adjusted delta, whereas an increase leaves it unchanged, such that Δ_{adj} penalises models whose performance improves at the expense of degraded general ability.

The **generalisation gap** (GG) measures how far OOD performance lags behind ID performance:

$$\text{GG} = S_{\text{OOD}} - S_{\text{ID}}, \quad (3)$$

where S_{OOD} and S_{ID} are the model’s respective scores. Large negative gaps indicate poor generalisation.

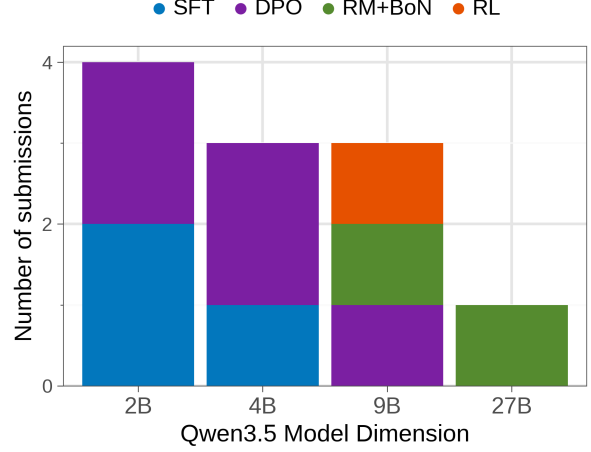


Figure 1: Number of submission by baseline model and improvement approach. Submission blending approaches are subsumed under what came last in their pipeline, e.g., SFT+DPO is counted as DPO.

The **proportion of error reduction** (PER) allows improvements from models with different baseline scores to be directly compared. Larger models generally start from stronger baselines and therefore have less room to improve, so an equal gain in score does not represent an equal reduction in the remaining error. PER divides the gain, Δ , by the difference between the baseline score, B , and the maximum score of 100:

$$\text{PER} = \frac{\Delta}{100 - B}. \quad (4)$$

It therefore expresses the gain as a proportion of the improvement still possible from that baseline.

We excluded Hot Air Balloon from the in-domain set of games because baseline models consistently hit the time wall.

4 Submissions

Challenge participants were asked to submit a paper through OpenReview describing their methodology, experimental setup, training data, and training and validation results. They were also required to submit a model to the Hugging Face Hub, with a model card containing the checkpoint, configuration files, and training card.

We received 11 model submissions from eight teams, six of which submitted papers to OpenReview: CityUoL, DAIR, playornotplay, Bentel rockers, Dialogue Architects, Large Language Problems. Although the challenge placed no restrictions on model choice, all submissions used models from the Qwen3.5 family. As reported in

Model	Team	ID Δ	OOD Δ	External Benchmarks	Adjusted ID Δ	Adjusted OOD Δ	Generalization Gap	Error Reduction (ID)	Error Reduction (OOD)
qwen3.5-2b		6.70	3.72	41.92	—	—	—	—	—
Qwen-GuidePlay-2B-v1	CityUoL 🐼	40.62	6.53	-2.06	38.56	4.47	-37.07	43.54%	6.78%
sft-dpo-v2	DAIR ★	46.42	10.84	-2.40	44.02	8.44	-38.56	49.75%	11.26%
sft-v1	DAIR	41.54	11.90	-4.55	36.99	7.35	-32.62	44.52%	12.36%
playornotplay-v1.0	playornotplay ✂	33.54	4.16	1.01	33.54	4.16	-32.36	35.95%	4.32%
qwen3.5-4b		29.28	17.99	53.53	—	—	—	—	—
Bentel_iter	Bentel rockers 🦖	4.02	-1.39	-3.21	0.81	-4.60	-16.70	5.68%	-1.69%
Bentel_iter_2	Bentel rockers	-0.81	-7.80	-9.89	-10.70	-17.69	-18.28	-1.15%	-9.51%
Bentel_iter_3	Bentel rockers	4.46	-1.39	-3.21	1.25	-4.60	-17.14	6.31%	-1.69%
qwen3.5-9b		37.37	24.91	53.85	—	—	—	—	—
prm-search-best_of_n	BSU-SLIM ▼	0.17	-9.73	-7.09	-6.92	-16.82	-22.36	0.27%	-12.96%
SCoRe_Qwen3.5-9B	Dialogue Architects ■	2.72	-1.54	0.48	2.72	-1.54	-16.72	4.34%	-2.05%
llp-final	Large Language Problems ▲	22.82	-2.82	0.64	22.82	-2.82	-38.10	36.44%	-3.76%
qwen3.5-27b		64.40	43.51	63.35	—	—	—	—	—
prm-guided-best_of_n	SLED-BSU ▼	-10.43	-13.49	0.86	-10.43	-13.49	-23.95	-29.30%	-23.88%

Table 1: Challenge results. Rows with a grey background report the absolute scores of the baseline models; for submitted models all columns report differences from the corresponding baseline model. External Benchmarks reports the static-game ClemScore. Icons in the Team column are the team markers used in the result figures. Bold values with a pale highlight mark the category leaders highlighted in Section 5.

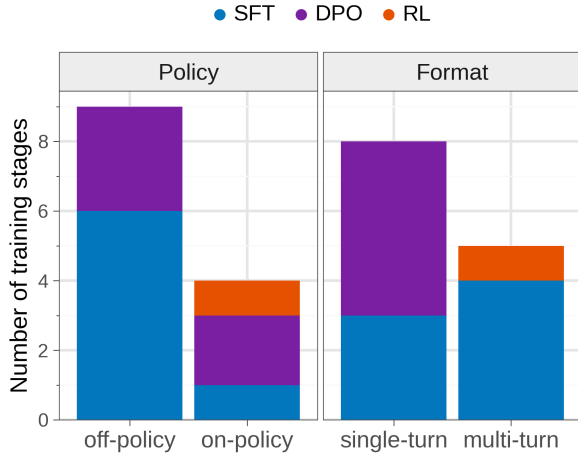


Figure 2: Policy and format of training strategies. Submissions blending approaches (e.g., SFT followed by DPO) are counted twice.

Fig. 1, the majority of submissions focused on improving smaller models, particularly the 2B version of Qwen3.5. The approaches included supervised fine-tuning (SFT), direct preference optimisation (DPO), on-policy reinforcement learning (RL), and an inference-time strategy combining a reward model with best-of- N sampling (RM+BoN). SFT was the most common approach, used by five of the six teams that submitted papers; Dialogue Architects was the exception. Four teams subsequently applied DPO. On-policy RL was used only in the final system submitted by Dialogue Architects; DAIR and playornotplay also reported experimenting with online RL but did not include it in their final submissions.

Figure 2 breaks down the training strategies

by data provenance (on-policy or off-policy) and interaction format (single-turn or multi-turn), following the categorisation in Horst et al. (2025). Most SFT approaches used off-policy data from Playpen, while DPO approaches were more evenly distributed across the two categories. All teams that used DPO experimented with single-turn data, whereas SFT was applied to both single-turn and multi-turn data. The Dialogue Architects submission was the only one to use online, on-policy training on multi-turn data.

5 Competition Results

Table 1 reports the challenge results for all submitted systems. DAIR’s 2B SFT+DPO system achieved the largest adjusted in-domain improvement (+44.02), in-domain error reduction (49.75%), and adjusted OOD improvement (+8.44). DAIR’s SFT-only variant achieved the largest OOD error reduction (12.36%). The largest improvement on the static benchmarks was achieved by playornotplay (+1.01).

6 Findings

Gains over the baseline are largest for the smallest model category. As shown in Figure 3, the largest absolute gains occur among the 2B systems, which also have the most room to improve relative to their baseline model. Results at larger model sizes are more mixed. The 4B and 9B systems preserve in-domain performance or achieve moderate to sizeable improvements, but often lose OOD performance. The 27B system, built on the

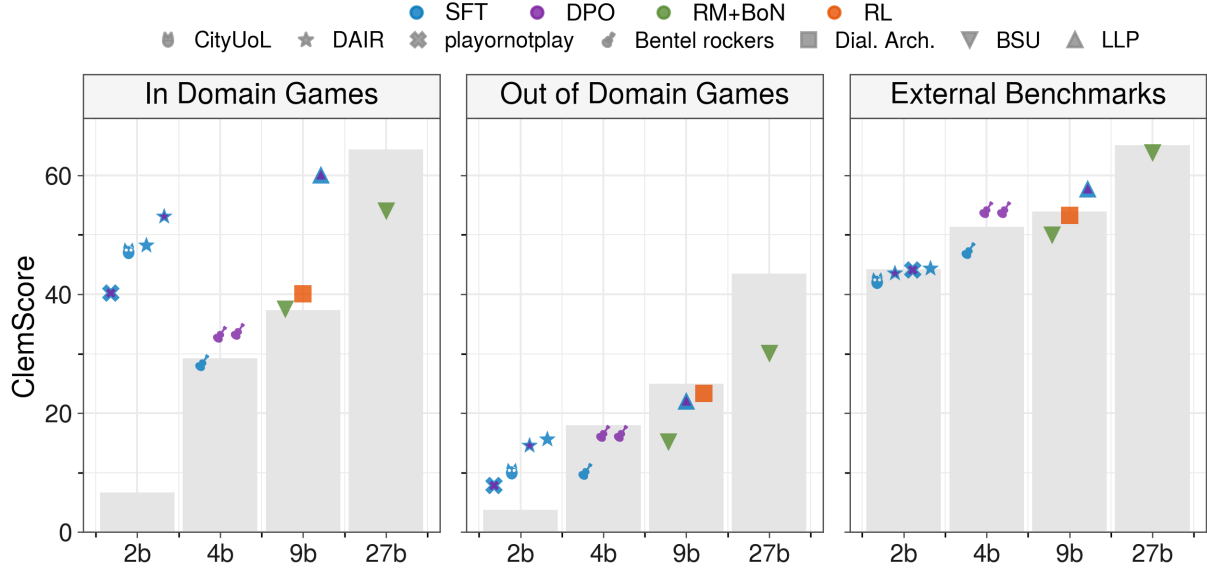


Figure 3: Results for training strategies: grey bars are the baseline model results, the coloured points are the results for each submission – divided in In Domain games, Out of Domain games, and External Benchmarks. Each submission is again categorised according to the training strategy. Teams identity is identifiable through marker icons.

strongest base model, regresses both with respect to in-domain and OOD performance.

Post-training changes the pattern of performance across games. Figure 6 shows that post-training does not simply raise or lower performance uniformly across the benchmark. At 2B, the SFT traces move outward most clearly on the TextMapWorld games and Deal or No Deal, while retaining the baseline model’s already high ImageGame score. Other game-specific weaknesses remain. At larger model sizes, post-trained variants likewise differ from their baseline models in which games improve, remain difficult, or decline.

OOD and static-benchmark gains are narrow and do not reliably track in-domain trends. Figure 3 shows that, at the aggregate level, in-domain improvements do not translate consistently into proportionate OOD gains. At the game level, OOD gains are concentrated in a subset of OOD games (particularly in Wordle variants) rather than spread across the full suite, as illustrated in Figure 6. Likewise, both aggregate and benchmark-level changes in static benchmark performance do not consistently track changes in interactive-game performance (see Figure 3 and Figure 6). Together, these results indicate a disconnect between aggregate improvements and a broader, game-independent ability gain.

Training with Dialogue Games Teaches to Follow Game Rules. The two metrics composing ClemScore measure different aspects of game play: % Played games is a proxy for the models’ instruction-following abilities in the game domain, while the Quality Score for models’ abilities in game playing. The aggregate results in Figure 4 show that the strongest 2B systems improve both these aspects. At larger model sizes, however, changes are less consistently aligned. In Domain for 4B, 9B and 27B submissions, game-level changes are mixed, with % Played games often improving where Quality Score tend to decrease with the exception of the LLP submission. Fig. 10, for the ID Taboo game, shows the mixed deltas in terms of quality score for a game that baseline Qwen models are able to play. Out of Domain, submissions tend to also complete more episodes, while their quality is lower. An example of an OOD game is Wordle Crazy in Fig. 11, where the improvements are mostly in terms of playability. In conclusion, training with dialogue games, with no emerging distinction among training strategies, makes models follow more strictly the structure and rules of the games. This is further confirmed by Fig. 9, showing that the most OOD quality improvements are in games that are structurally similar to ID ones (Wordle variants).

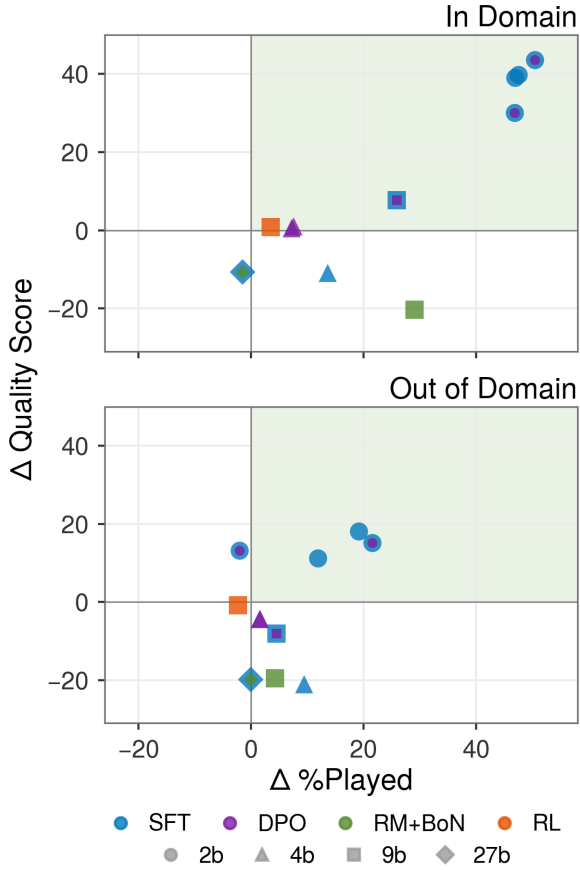


Figure 4: Submissions’ Δ % Played instances and Δ Quality Score for In Domain and Out of Domain Games, classified according training method and baseline size. 2B submissions tend to improve both these aspects, with the exception of playornotplay OOD, where unexpectedly it only improves quality score. In general, SFT + DPO seems to produce consistent results for ID games. For larger baselines, models tend to only increase the % of played game, with mixed performances ID and decreased in OD.

Dialogue Games Training Makes Language More Concise. Post-training with Dialogue Game improves following games rules and structure, does this benefit performances on external benchmarks? Table 2 shows that gains are obtained for BBH, with decreasing performances for IFEval in particular (replicating Horst et al., 2025). We observe that trained models tend to generate more concise answers, an interesting exception for this is the RL/SCoRE method, for which the output length is not significantly shorter on average. In BBH more concise outputs improve performances, since baseline models’ outputs suffer from formatting errors (i.e., unexpected chain-of-thought and complete repetition of the answer), that are avoided

by training with game language.⁵ As shown in Fig. 5, however, these improvements are balanced by decreases in terms of corrected answers on the remaining questions.

Policy and Dialogue Format. Following the distinction in Horst et al. (2025), Fig. 7 reports the influence of a series of factors on the final deltas. In particular, we consider: 1) the training regime: token vs turn loss, online vs offline training; and 2) the type of data used for training: off-policy vs on-policy, single-turn vs multi-turn. These patterns should be interpreted as trends rather than conclusive results. Nevertheless, off-policy data and multi-turn formats appear to yield consistent improvements in games, both in and out of domain. Learning off-policy from multi-turn dialogues means learning from complete demonstrations generated by stronger models, which may provide information about the overall game structure rather than a meaningful turn-level learning signal. Regarding online versus offline learning, the only online approach, SCoRE/RL, is notably conservative, as already shown in Fig. 4 and Table 2.

7 Conclusions

In this technical report, we discuss the outcome of the first LM Playschool 2026 challenge. The main goal of this challenge is to provide a controllable environment based on dialogue games for simulating situated language use and for studying how models learn from and through interaction, instead of passively observing dialogue data. To evaluate whether the resulting models are well-rounded language users, we first use unseen instances of dialogue games that the model has seen at training time. However, to make sure that models develop language abilities that do not exclusively apply to training data, our experimental setup values different forms of generalisation by evaluating: 1) performance on out-of-domain dialogue games, and 2) performance on reference-based NLP benchmarks.

The challenge received great interest with 11 submissions across a variety of different model sizes ranging from 2B up to 27B. Based on the categorisation of learning-from-interaction approaches proposed by Horst et al. (2025), most approaches focused on off-policy training regimes, while only

⁵This finding is accentuated by the BBH evaluation prompt, which does not specify the answer format and results in more errors by the baselines models, due to verbosity.

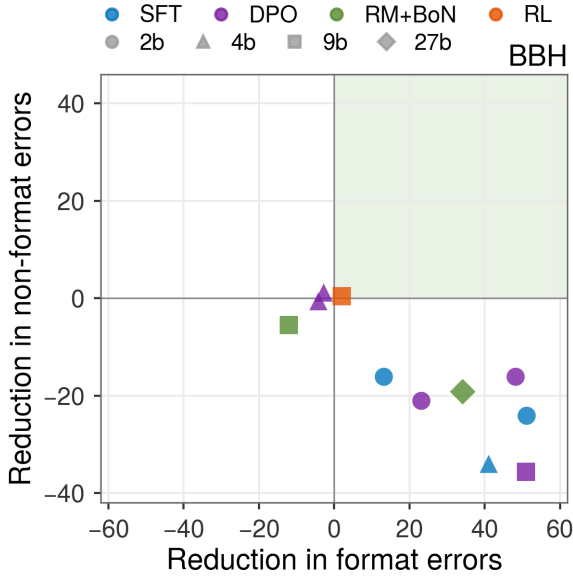


Figure 5: BBH error composition relative to the base model. For BBH questions ($n = 104$), every incorrect episode is classified as: a “format error”, where the error is due to a format issues (the answer is reached but buried in reasoning or prefixed, e.g. A: (C)), or “non-format error” for the remaining cases. Both axes are differences in instances counts, computed as reduction in errors (base – submission), so a positive value means the submission made fewer errors of that class than its baseline model; the shaded upper-right quadrant therefore marks a reduction in both classes at once. Marker shape denotes the base model, colour the terminal stage of the fine-tuning pipeline.

one participant used an on-policy training regime with Reinforcement Learning. Based on our results, current learning methods provide substantial gains, particularly for small language models, with their performance matching or sometimes exceeding base model performance on in-domain games. However, generalisation remains limited, suggesting that models tend to learn from the surface form of the language outputs they observe.

While the submissions demonstrate that post-training on recorded Playpen interactions can produce substantial improvements, these improvements are uneven across interactive games. The feedback taxonomy also identifies three promising directions that the challenge did not test: training directly on within-episode Game Master corrections, refreshing preference data as the policy improves, and learning from feedback supplied—implicitly or explicitly—by a co-player. Evaluating these regimes would extend the challenge from learning game protocols from recorded trajectories towards learning to adapt to feedback during interaction.

Challenge submissions did not experiment with chain-of-thought reasoning, and training made the models’ language more concise rather than more strategic. Related works have demonstrated that training with chain of thought on competitive (Liu et al., 2026) and cooperative (Yuan et al., 2026) games can teach models more strategic chain-of-thought reasoning, leading to improvements on general benchmarks. Given the number of games and high-level abilities captured by the Playpen game environment, incorporating reasoning traces into training could lead to valuable improvement.

Modelling the multi-turn nature of dialogue games is essential to perform correct state-tracking across turns. We believe that the Playpen environment will facilitate the implementation of multi-turn training regimes that can leverage signals that can be derived across multiple turns of interaction.

Finally, building on previous work on multi-modal dialogue games (Hakimov et al., 2025), studying the intersection between communicative grounding and visual grounding represents a promising research direction for future multimodal conversational agents (Suglia et al., 2024).

Acknowledgements

We thank Uday Bhaskar, Kranti Chalamalasetti, Shahrukh Mohiuddin, Karl Osswald, and Ruilin Yang for their contributions to the out-of-distribution games. We gratefully acknowledge the support of the Deutsche Forschungsgemeinschaft (DFG), the Autonomous Province of Bozen–Bolzano, and Amazon. This work was conducted as part of the LLADIGA project under the Joint Projects Italia–Germania 2025 programme. Finally, we thank ELLIS for bringing together the organisers and facilitating this collaboration.

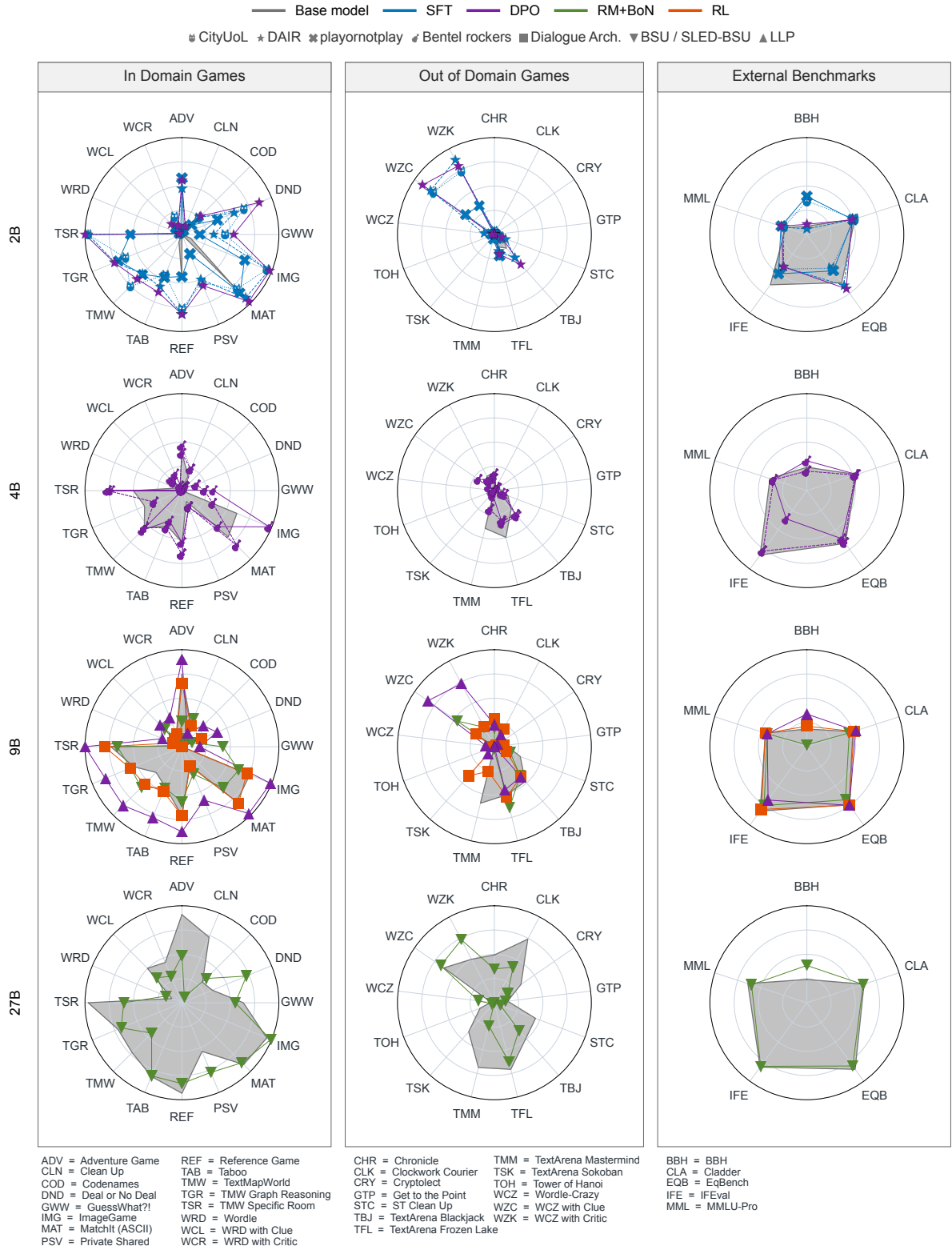


Figure 6: Per-game ClemScore profiles for every base-model size (2B, 4B, 9B, 27B) and its submitted post-trained variants. Each spoke denotes a game or benchmark, and radial distance gives the score (0–100). The grey trajectory is the corresponding base model; coloured trajectories are its post-trained variants. Marker icons identify the submitting team, matching Table 1; colours identify the final-stage method. Differences from the grey line show which game-specific strengths and weaknesses post-training preserved or changed.

A Detailed results

The following plots complement the aggregate and profile analyses in Section 6. Table 2 illustrates how individual benchmark-level changes contribute to an aggregate increase or decrease in external benchmark performance. The training-dimension map in Figure 7 situates submissions in the feedback taxonomy, providing descriptive diagnostics rather than controlled estimates of a training-design effect. The radar plots in Figures 8 and 9 unpack the ClemScore profiles presented in Figure 6 into their two component parts: outward movement against the base-model in the Quality Score profiles indicates better play on that game, whereas outward movement in the percentage-played profiles indicates that more episodes were completed. Comparing the two identifies whether a game-level change reflects quality, completion, or both. Finally, the plots in Figures 10 and 11 isolate concrete example cases where post-training flipped a base model’s behaviour on an instance-level.

Model	Quality Score					Mean words generated				
	BBH	CLad.	EQ-B.	IFE.	MMLU	BBH	CLad.	EQ-B.	IFE.	MMLU
qwen3.5-2b	8.65	52.00	63.20	64.00	23.00	196.1	483.3	8.0	611.3	6.4
GuidePlay-2B	25.97	0.00	-15.00	-20.00	2.00	1.5	1.2	8.0	346.4	4.0
sft-dpo-v2	1.93	-3.00	5.84	-23.00	5.00	68.2	73.7	8.0	228.0	4.0
sft-v1	-2.88	-2.00	2.23	-23.00	2.00	77.5	141.8	8.0	577.5	4.0
playornotplay	30.77	-2.00	-11.11	-14.00	5.00	84.6	220.6	8.0	433.7	4.0
qwen3.5-4b	24.04	54.00	68.68	82.00	40.00	150.2	612.8	26.2	193.7	4.0
Bentel_iter	-3.85	-2.00	-1.28	-5.00	-4.00	220.4	618.9	8.9	370.7	4.0
Bentel_iter_2	6.73	0.00	-6.79	-46.00	-4.00	25.9	93.2	9.5	261.5	4.0
Bentel_iter_3	-3.85	-2.00	-1.28	-5.00	-4.00	220.4	618.9	8.9	370.7	4.0
qwen3.5-9b	18.27	50.00	73.99	82.00	45.00	179.7	702.7	8.0	191.0	4.0
prm-search-bon	-16.35	-4.00	-5.59	-7.00	-1.57	224.7	360.3	20.0	196.6	5.2
SCoRe-9B	2.88	1.00	0.51	-2.00	0.00	172.0	718.3	8.0	182.1	4.0
llp-final	15.38	3.00	1.47	-14.00	-1.57	15.4	180.3	12.4	168.5	4.0
qwen3.5-27b	24.04	62.00	84.72	82.00	64.00	191.7	755.3	8.0	193.4	4.0
prm-guided-bon	14.42	-1.00	-4.14	-1.00	-4.00	61.3	299.9	8.0	178.4	4.0

Table 2: External benchmark results deltas (Quality Score), with the mean number of words generated per question (grey columns). Score columns show absolute values for base models and differences to the base for submissions; length columns are absolute in both cases. **Bold** marks a significant reduction in output length relative to the base model (Wilcoxon signed-rank over paired questions, $p < 0.05$).

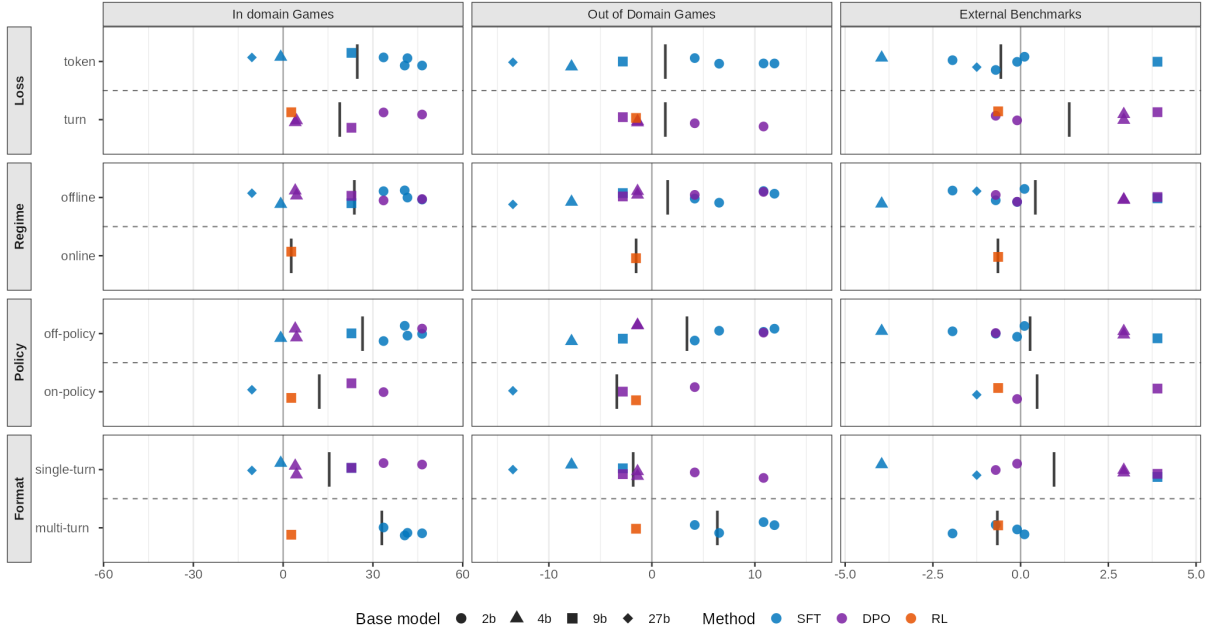


Figure 7: Submission-level map of score changes by four annotated training dimensions: loss granularity, training regime, policy provenance, and dialogue format. Each coloured point is a post-trained submission’s change from its base model; colours denote method and shapes denote base-model size. Black markers show the corresponding base-model results. A submission can appear in more than one category when its pipeline combines training stages. This figure is descriptive: the categories overlap and are not controlled comparisons of the effects of individual design choices.

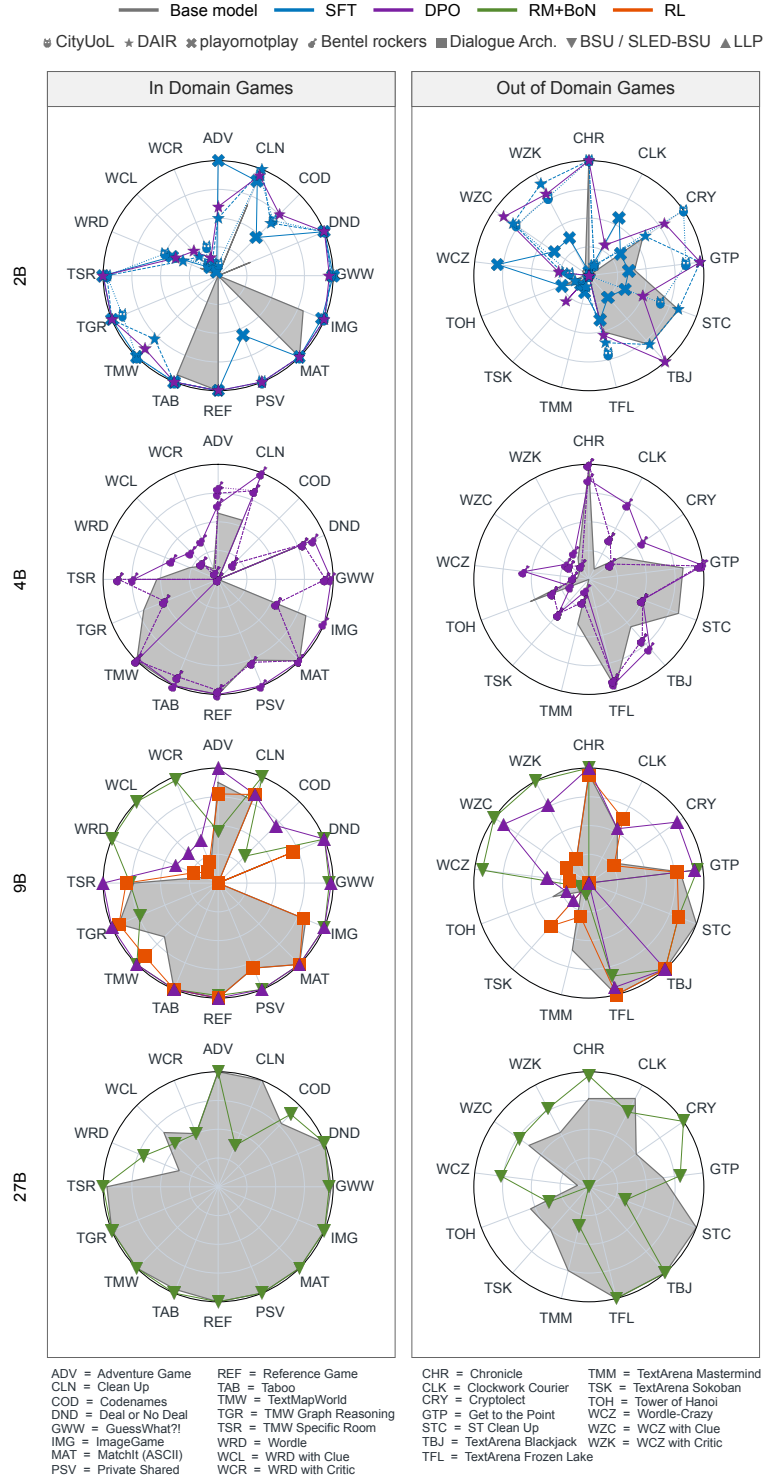


Figure 8: Per-game percentage-played profiles for every base-model size (2B, 4B, 9B, 27B) and its submitted post-trained variants. This decomposes the ClemScore profiles shown in Figure 6. Each spoke denotes a dynamic game, and radial distance gives the percentage of episodes completed (0–100). The grey trajectory is the corresponding base model; coloured trajectories are its post-trained variants. Marker icons identify the submitting team, matching Table 1; colours identify the final-stage method. Differences from the grey line show which game-specific completion strengths and weaknesses post-training preserved or changed. Static benchmarks are omitted because completion is near-saturated and does not materially aid interpretation.

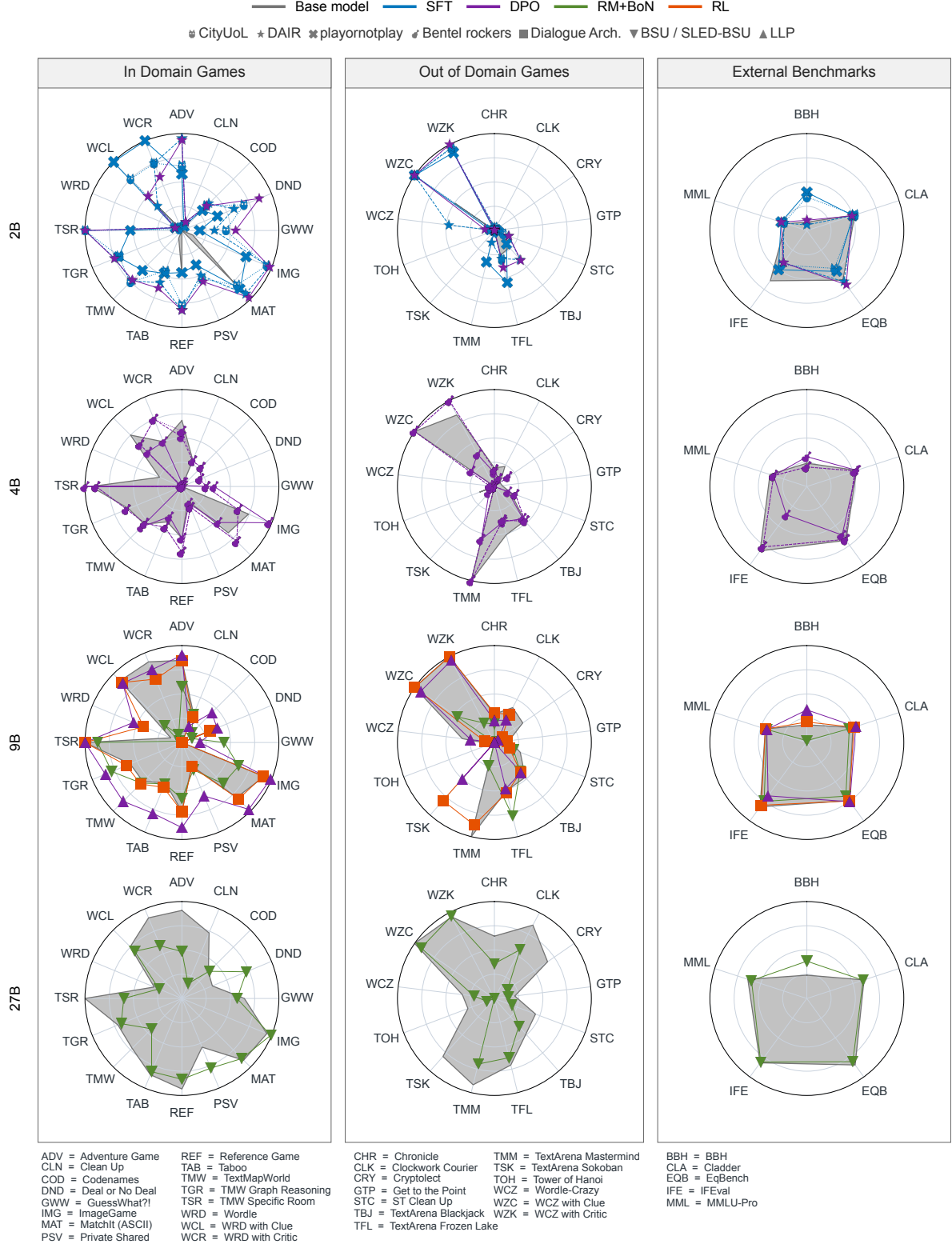


Figure 9: Per-game Quality Score profiles for every base-model size (2B, 4B, 9B, 27B) and its submitted post-trained variants. This decomposes the ClemScore profiles shown in Figure 6. Each spoke denotes a game or benchmark, and radial distance gives mean Quality Score (0–100). The grey trajectory is the corresponding base model; coloured trajectories are its post-trained variants. Marker icons identify the submitting team, matching Table 1; colours identify the final-stage method. Differences from the grey line show which game-specific strengths and weaknesses post-training preserved or changed.

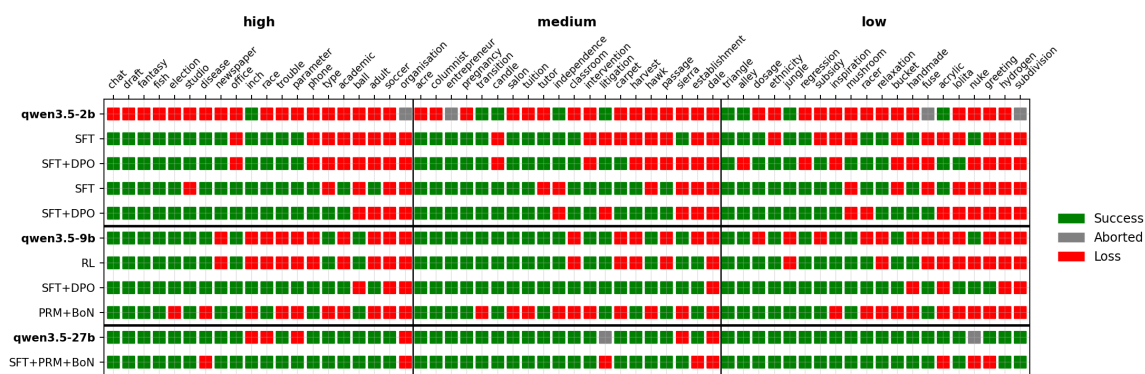


Figure 10: ID Taboo instances flipped from the base models to the trained models. The Qwen baselines are able to play the game, with very few aborted instances even for the smallest 2B baseline. Trained submissions lead to improvements in terms of quality scores (flipped instances).

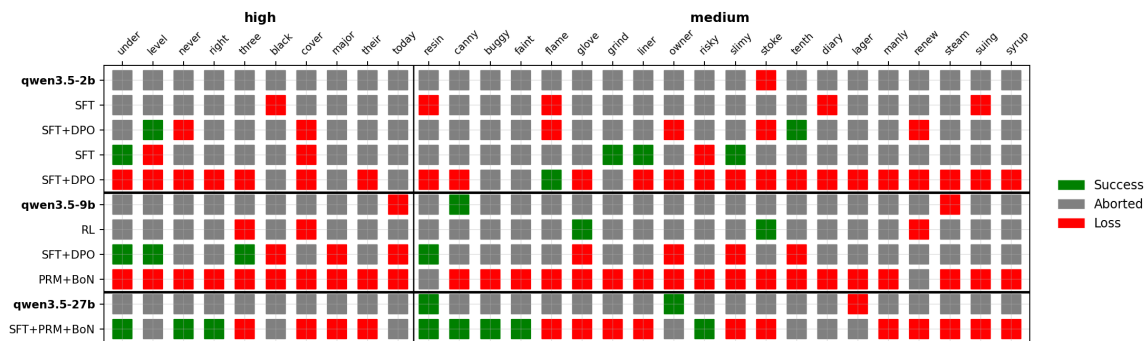


Figure 11: OOD Wordle Crazy instances flipped from the base models to the trained models. Qwen baselines tend to abort all episodes, trained submissions are able to follow the rules of the game and play, but mostly losing it (reflected in the quality score).

References

- Emily M Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5185–5198. Issue: 2.
- Anne Beyer, Kranti Chalamalasetti, Sherzod Hakimov, Brielen Madureira, Philipp Sadler, and David Schlangen. 2024. [clembench-2024: A Challenging, Dynamic, Complementary, Multilingual Benchmark and Underlying Flexible Framework for LLMs as Multi-Action Agents](#). *arXiv preprint*. ArXiv:2405.20859 [cs].
- Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. [Experience grounds language](#). *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 8718–8735. ArXiv: 2004.10151 ISBN: 9781952148606.
- Paul Bloom. 2000. *How children learn the meanings of words*. Learning, development, and conceptual change. MIT Press, Cambridge.
- Kranti Chalamalasetti, Jana Götze, Sherzod Hakimov, Brielen Madureira, Philipp Sadler, and David Schlangen. 2023. [clembench: Using game play to evaluate chat-optimized language models as conversational agents](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11174–11219, Singapore. Association for Computational Linguistics.
- Pengyu Cheng, Tianhao Hu, Han Xu, Zhisong Zhang, Yong Dai, Lei Han, Nan Du, and Xiaolong Li. 2024. [Self-playing Adversarial Language Game Enhances LLM Reasoning](#).
- Eve V. Clark. 2016. *First language acquisition*, 3rd ed edition. Cambridge University Press, Cambridge.
- Alejandrina Cristia, Emmanuel Dupoux, Michael Gurven, and Jonathan Stieglitz. 2019. Child-directed speech is infrequent in a forager-farmer population: A time allocation study. *Child development*, 90(3):759–773.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Leon Guertler, Bobby Cheng, Simon Yu, Bo Liu, Leshem Choshen, and Cheston Tan. 2025. [Textarena](#). *Preprint*, arXiv:2504.11442.
- Mustafa Omer Gul and Yoav Artzi. 2024. [CoGen: Learning from Feedback with Coupled Comprehension and Generation](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12966–12982, Miami, Florida, USA. Association for Computational Linguistics.
- Sherzod Hakimov, Yerkezhan Abdullayeva, Kushal Koshti, Antonia Schmidt, Yan Weiser, Anne Beyer, and David Schlangen. 2025. [Using game play to investigate multimodal and conversational grounding in large multimodal models](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5686–5718, Abu Dhabi, UAE. Association for Computational Linguistics.
- Betty Hart and Todd R. Risley. 1995. *Meaningful differences in the everyday experience of young American children*. Paul H. Brookes Publishing Co.
- Nicola Horst, Davide Mazzaccara, Antonia Schmidt, Michael Sullivan, Filippo Momentè, Luca Franceschetti, Philipp Sadler, Sherzod Hakimov, Alberto Testoni, Raffaella Bernardi, Raquel Fernández, Alexander Koller, Oliver Lemon, David Schlangen, Mario Giulianelli, and Alessandro Suglia. 2025. [Playpen: An environment for exploring learning from dialogue game feedback](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29854–29891, Suzhou, China. Association for Computational Linguistics.
- Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. 2024. [Cladder: Assessing causal reasoning in language models](#). *Preprint*, arXiv:2312.04350.
- Jiatong Li, Rui Li, and Qi Liu. 2023. Beyond Static Datasets: A Deep Interaction Approach to LLM Evaluation. ArXiv: 2309.04369v1.
- Tal Linzen. 2020. [How can we accelerate progress towards human-like linguistic generalization?](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics.
- Bo Liu, Simon Yu, Zichen Liu, Leon Guertler, Penghui Qi, Daniel Balcells, Mickel Liu, Cheston Tan, Weiyan Shi, Min Lin, Wee Sun Lee, and Natasha Jaques. 2026. [SPIRAL: Self-play on zero-sum games incentivizes reasoning via multi-agent multi-turn reinforcement learning](#). In *The Fourteenth International Conference on Learning Representations*.
- Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in cognitive sciences*.
- Filippo Momentè, Alessandro Suglia, Mario Giulianelli, Ambra Ferrari, Alexander Koller, Oliver Lemon,

- David Schlangen, Raquel Fernández, and Raffaella Bernardi. 2025. [Triangulating LLM progress through benchmarks, games, and cognitive tests](#). *arXiv preprint arXiv:2502.14359*.
- OpenEnv Technical Committee. 2025. OpenEnv: Agentic Execution Environments for Reinforcement Learning and Post-Training. <https://github.com/huggingface/OpenEnv>.
- Samuel J. Paech. 2023. [Eq-bench: An emotional intelligence benchmark for large language models](#). *Preprint*, arXiv:2312.06281.
- Dan Qiao, Chenfei Wu, Yaobo Liang, Juntao Li, and Nan Duan. 2023. [GameEval: Evaluating LLMs on Conversational Games](#). ArXiv: 2308.10032.
- Qwen Team. 2026. [Qwen3.5: Towards native multi-modal agents](#).
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*, pages 53728–53741, Red Hook, NY, USA. Curran Associates Inc.
- David Schlangen. 2023. [Dialogue games for benchmarking language understanding: Motivation, taxonomy, strategy](#). *CoRR*, abs/2304.07007.
- David Schlangen, Sherzod Hakimov, Chalamalasetti Kranti, Jonathan Jordan, and Philipp Sadler. 2026. [A third paradigm for llm evaluation: Dialogue game-based evaluation using clembench](#). *Preprint*, arXiv:2507.08491.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazary, and 431 others. 2023. [Beyond the imitation game: Quantifying and extrapolating the capabilities of language models](#). *Transactions on Machine Learning Research*.
- Alessandro Suglia, Ioannis Konostas, and Oliver Lemon. 2024. [Visually Grounded Language Learning: a review of language games, datasets, tasks, and models](#). *Journal of Artificial Intelligence Research*, 79:173–239. ArXiv: 2312.02431.
- Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. [Challenging BIG-bench tasks and whether chain-of-thought can solve them](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.
- J Terry, Benjamin Black, Nathaniel Grammel, Mario Jayakumar, Ananth Hari, Ryan Sullivan, Luis S Santos, Clemens Dieffendahl, Caroline Horsch, Rodrigo Perez-Vicente, et al. 2021. [Pettingzoo: Gym for multi-agent reinforcement learning](#). *Advances in Neural Information Processing Systems*, 34:15032–15043.
- Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, Rodrigo Perez-Vicente, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. 2025. [Gymnasium: A standard interface for reinforcement learning environments](#). *Preprint*, arXiv:2407.17032.
- Ruiyi Wang, Haoifei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Yonatan Bisk, Graham Neubig, and Hao Zhu. 2024a. [SOTOPIA-pi: Interactive Learning of Socially Intelligent Language Agents](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12912–12940, Bangkok, Thailand. Association for Computational Linguistics.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhui Chen. 2024b. [Mmlu-pro: a more robust and challenging multi-task language understanding benchmark](#). In *Proceedings of the 38th International Conference on Neural Information Processing Systems, NIPS '24*, Red Hook, NY, USA. Curran Associates Inc.
- Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023. [An Empirical Study on Werewolf](#). ArXiv: 2309.04658v1.
- Huining Yuan, Zelai Xu, Zheyue Tan, Xiangmin Yi, Mo Guang, Kaiwen Long, Haojia Hui, Boxun Li, Xinlei Chen, Bo Zhao, Xiao-Ping Zhang, Chao Yu, and Yu Wang. 2026. [MARSHAL: Incentivizing multi-agent reasoning via self-play with strategic LLMs](#). In *The Fourteenth International Conference on Learning Representations*.
- Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. [Instruction-following evaluation for large language models](#). *Preprint*, arXiv:2311.07911.