PLAYPEN: An Environment for Exploring Learning From Dialogue Game Feedback

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**[ᵡ], **Alessandro Suglia**[◇]
[⊗]University of Potsdam, [‡]University of Trento, [§]Saarland University, [◁]ETH Zurich, [#]Amsterdam UMC, [⋆]Free University of Bozen Bolzano, [△]University of Amsterdam, [°]Heriot-Watt University, [⊕]DFKI, [ᵡ]UCL, [◇]University of Edinburgh
lm-playschool@googlegroups.com

Abstract

Interaction between learner and feedback-giver has come into focus recently for post-training of Large Language Models (LLMs), through the use of reward models that judge the appropriateness of a model's response. In this paper, we investigate whether Dialogue Games—goal-directed and rule-governed activities driven predominantly by verbal actions—can also serve as a source of feedback signals for learning. We introduce PLAYPEN, an environment for off- and online learning through Dialogue Game self-play, and investigate a representative set of post-training methods: supervised fine-tuning; direct alignment (DPO); and reinforcement learning with Group Relative Policy Optimization (GRPO). We experiment with post-training a small LLM (Llama-3.1-8B-Instruct), evaluating performance on unseen instances of training games as well as unseen games, and on standard benchmarks. We find that imitation learning through SFT improves performance on unseen instances, but negatively impacts other skills, while interactive learning with GRPO shows balanced improvements without loss of skills. We release the framework and the baseline training setups to foster research in this promising new direction of "learning in (synthetic) interaction".

https://github.com/lm-playpen/playpen

1 Introduction

Dialogue Games of the kind illustrated in Figure 1 have been used to evaluate interactive reasoning capabilities of LLMs, requiring as they do the strategic acquisition of information through multi-turn interaction (Bertolazzi et al., 2023; Chalamalasetti et al., 2023; Qiao et al., 2023; Li et al., 2023; Guertler et al., 2025; Cui et al., 2025). In this paper, we show that it is also possible to derive a learning signal from such games, which we call Dialogue Game Feedback (DGF). DGF offers a unique combination of properties: A) Unlike feedback from the reward models used in Reinforcement Learning from Human Feedback (RLHF; Ouyang et al., 2022), DGF is *teleological* or goal-directed, in that it judges not just the appropriateness of a response, but also its effectiveness in advancing toward a desired outcome. This is a property DGF shares with feedback from process- and outcome-based reward models used to optimize reasoning models (PRM, Setlur et al. 2024; ORM, Hosseini et al. 2024; Cobbe et al. 2021; respectively). B) Unlike these aforementioned prior methods, DGF is *objective*, in that it can be computed programmatically, rather than needing a learned, "subjective" judgement model; a property it shares with the "verifiable rewards" of Lambert et al. (2025). C) Unlike all of these other methods, it can be derived from *inter-subjective*, multi-turn

A (quietly reads game card): `Target word: ugly. Taboo words: displeasing, despicable, unattractive.`
(1a) A (to B): Something that is not pleasing to the eye, often making a bad impression.
(1b) B: Eyesore
(2a) A: More about the quality of a person or thing, rather than a nuisance or distraction.
(2b) B: Ugly.
(3a) A: Yay! Well done us.

Figure 1: A successful episode of the two-player Dialogue Game `Taboo`, as played by humans.

[*]Joint first authorship.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 29843–29880
November 4-9, 2025

linguistic interaction. As DGF is defined in terms of *task success*, and task success here is conditional on *communicative success*—i.e., players are required to produce mutually intelligible language—the feedback signal implicitly carries linguistic normative pressure. D) Lastly, where other methods focus on alignment with respect to *desirability* (RLHF) or specifically on reasoning skills in domains such as maths and coding (PRM, ORM, verifiable rewards), DGF rewards general backwards- and forwards-looking language use. At the same time, insofar as the underlying game requires them, DGF also targets specific skills such as spatial reasoning or referential language.

Figure 1 illustrates this. It shows an example of two players playing the Dialogue Game `Taboo` where a *clue giver* needs to describe a concept to a *guesser*, while avoiding certain "taboo" words. Both players produce natural language strings, which can be judged in two ways: *teleologically*, where both 1a and 1b are appropriate but do not lead to success, while 2a and 2b succeed; and *objectively* as it can be programmatically verified that 1b and 2b comply with the rules and that 2b is the correct answer. Moreover, for the game to progress, both clue giver and guesser must produce mutually intelligible language, placed in the context of the interaction as a whole. E.g. turn 2a must consider the previous guess and how it failed (backwards-looking) in order to produce an utterance that aims to elicit a better guess (forward-looking).

Contribution 1: We introduce PLAYPEN, an environment where LLMs engage in dialogue games (without human intervention) from which a feedback signal for learning is derived; either online during gameplay or offline from rollouts, and based either on single trajectories or sets of ranked alternatives. **Contribution 2:** We demonstrate how to leverage this feedback signal for post-training.[1] We investigate three learning approaches: imitation learning (supervised fine-tuning); an offline alignment approach, Direct Preference Optimization (DPO; Rafailov et al., 2023); and an online learning algorithm, Group Relative Policy Optimization (GRPO; Shao et al., 2024)—thereby establishing strong baselines for this learning environment. We evaluate the resulting models on a range of tests, including held-out dialogue games to assess skill generalisation, a comprehensive suite of tests assessing broader linguistic and cognitive abilities (Momentè et al., 2025), and standard NLP benchmarks such as MMLU-Pro (Wang et al., 2024b), Big-bench Hard (Suzgun et al., 2023), and IFEval (Zhou et al., 2023b). **Contribution 3:** We show that imitation learning through SFT improves performance on unseen instances but negatively impacts other skills, whereas interactive learning with GRPO achieves balanced improvements without skill degradation.

2 Related Work

Our work builds on several threads of research, which we briefly review here and visualise in Fig. 2.

Sources of Learning Feedback. Post-training methods typically assume some *feedback* on the appropriateness of a model's production R in a context C, via a feedback function f. We outlined in the introduction the main approaches to defining f, from trained judge models to verifiers in formal domains (see Kumar et al., 2025; Lambert, 2024, for recent overviews), and positioned our *Dialogue Game Feedback* within this landscape. The feedback function either judges (C, R)—we may call this *direct feedback*—or ranks candidate responses $(C, \{R, R', \dots\})$ in a *comparative judgement*. Most methods judge the *whole* production R; but see Wu et al. (2023) for more fine-grained proposals. We can make one further distinction. The approaches discussed so far can be seen as providing *outside* or *3rd person feedback*, in that the judge observes but does not participate in the interaction. Others (e.g., Sumers et al., 2021; Chen et al., 2024) extract evidence from the ongoing interaction itself, or inject such feedback synthetically (Akyurek et al., 2023; Wang et al., 2024a)—what we call *inside* or *1st person* feedback.[2] It has been variously explored (see e.g. Sumers et al. 2021; Chen et al. 2024) to take evidence from the ongoing interaction as well. Our framework can incorporate such "inside" feedback, but here we focus on what can be learned from "outside"—programmatically computed game feedback. We regard DGF as both *teleological* (goal-directed) and *objective*, in that f is defined algorithmically and encodes the game's specific goals. A similar kind of feedback has been used recently by Gul and Artzi (2024), but only in

[1] Leaving to future work the exploration of these types of interactive settings for language acquisition from scratch.

[2] There is a much older parallel line on feedback in language learning, from Luc Steels' work (Steels, 2003; Nevens and Spranger, 2017) to more recent 'emergent communication' studies (e.g., Cao et al., 2018). How to adapt this for LLM post-training remains open.

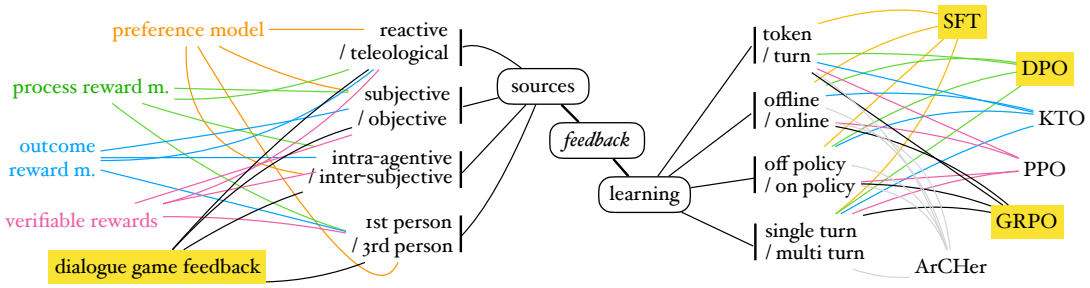

Figure 2: Placing Dialogue Game Feedback in the wider research context. Highlighted on the right the selection of learning methods that we provide baselines for.

the context of a single game, and with specialised learning methods. Similarly, Sadler et al. (2024) used feedback from a cognitively inspired programmatic partner to bootstrap the language capabilities of a collaborative neural agent in a single visual reference task.

Methods for Learning from Feedback. Assuming that a feedback function f is at hand, how can this be used to improve the model's production policy? Various methods have been developed in recent years to do this (see the surveys cited above). A first way of categorising them is according to their granularity. *Token-based* methods, such as supervised fine-tuning, essentially use f to filter data for imitation learning. The hypothesis is that producing sequences like those observed during training generalises to improving the model's strategic behaviour. *Turn-level* methods instead make direct use of the turn-level feedback obtained from f. Within these, *offline* variants such as Direct Preference Optimization (DPO; Rafailov et al., 2023) and Kahneman-Tversky Optimization (KTO; Ethayarajh et al., 2024) train on preference judgements from previously collected data, which do not need to come from the current policy (in this sense, they are both offline and off-policy). On the other hand, *online* methods such as Proximal Policy Optimization (PPO; Schulman et al., 2017) and Group Relative Policy Optimization (GRPO; Shao et al., 2024) attempt to directly improve the policy that produces the samples. Finally, multi-turn methods such as ArCHer (Zhou et al., 2024b), ReSpect (Chen et al., 2024), and REFUEL (Gao et al., 2024) are emerging, which explicitly handle the turn-level structure of conversations; we leave these for future work and here focus on a representative selection of methods: SFT, DPO, and GRPO.

LLMs and Dialogue Games. Conversational interactions framed as games have long been used to investigate language use; see discussion in (Schlangen, 2019, 2023; Suglia et al., 2024), which also proposed to use them for evaluating language use capabilities of NLP models. This idea has been implemented by various frameworks in recent years. Early frameworks such as TextWorld (Côté et al., 2019) supported only a single game genre (interactive fiction) and assumed specialised models. Only with the advent of generalist models that can be *prompted* into being specialists (Brown et al., 2020; Wei et al., 2021) did it become possible to implement this idea at a larger scale, both for single games (Bertolazzi et al., 2023; Mazzaccara et al., 2024) and in multi-game frameworks (Chalamalasetti et al., 2023; Qiao et al., 2023; Li et al., 2023; Gong et al., 2023; Wu et al., 2024; Zhou et al., 2024a; Duan et al., 2024; Guertler et al., 2025; Cui et al., 2025). Among these, we chose to build on `clembench` (Chalamalasetti et al., 2023) as the longest-running continuously maintained effort, which also comes with an extensive archive of dialogue game transcripts spanning a wide range of both open and closed models.

The Cognitive Plausibility of Learning from Interaction. It is well established in the developmental literature that human language acquisition requires social interaction (Clark, 2016; Bruner and Watson, 1983; Kuhl, 2007; Hiller and Fernández, 2016; Saxton, 2000; Bloom, 2000), and similar arguments have been made for machine language learning (Fernández et al., 2011; Bisk et al., 2020; Bender and Koller, 2020), especially given stark differences in sample efficiency between human and artificial learners (Hart and Risley, 1995; Cristia et al., 2019; Linzen, 2020). Our work speaks to this hypothesis by exploring a learning signal derived from (an approximation of) social linguistic interaction. Incidentally, this is also the motivation of the BabyLM challenge (Charpentier et al., 2025),

which in its latest incarnation explicitly encourages the use of synthetic interaction.

A few recent pioneering works (Nikolaus and Fourtassi, 2021; Ma et al., 2025) have begun to explore this direction, demonstrating the potential benefits of interaction even in learning from scratch. In this work, our focus is on post-training, and we assume that models can already follow instructions well enough to engage in gameplay.

3 Playpen: Dialogue Games & Feedback

3.1 Dialogue Games with LLMs

Figure 1 above provided an example of a Dialogue Game. How can we enable LLMs to play such games effectively? One of the surprising insights of the "LLM revolution" was that, at previously unseen scale, these models can be *prompted* to perform a wide range of tasks (Wei et al., 2021; Brown et al., 2020). As the frameworks described above have shown, this extends to prompting LLMs to act as policies for playing conversational games—albeit with varying degrees of success. For example, a simple prompt such as "*We are playing a word guessing game. Your task is to describe the word, but you are not allowed to use some other words. The word to describe is 'ugly', and the words to avoid are 'displeasing', 'despicable', and 'unattractive'.*" can induce (at least some) LLMs to act as a policy π_{taboo} capable of playing that specific role in the game reasonably well.[3] To enable self-play with LLMs, each player must be separately prompted, often with distinct information states. Following Chalamalasetti et al. (2023); Smith et al. (2024), we use a programmatic *Game Master* (GM) to mediate the interaction. In the case of Figure 1, for example, the GM would insert a turn between 1a and 1b, delivering instructions to player B and relaying the clue from player A. See Appendix C for full transcripts of such interactions.

All of the games used here (see Section 3.3 below) involve some form of reasoning. Crucially, however, the reasoning involved is fundamentally different from that required in standard applications of reasoning models (Besta et al., 2025). Unlike conventional reasoning tasks such as math word problems (Hendrycks et al., 2021), which are *well-posed*—i.e., the problem is fully specified and the challenge lies in deriving a solution through a correct sequence of steps—the games studied here require *multi-turn* and *interactive* reasoning, as they are *ill-posed* at the outset. They only become tractable through iterative exchanges between players. Consider the starting prompt for a `Wordle`-type word guessing game: "*Guess a 5 letter word*". Only through making guesses, receiving feedback, and updating beliefs accordingly does the task become solvable and the identity of the target word recoverable. The reasoning at play in such settings involves managing uncertainty and coordinating with another agent under conditions of imperfect information.

3.2 Dialogue Game Feedback

We refer to games such as `Taboo` or `Wordle` as *dialogue games* (DG). A particular instantiation of a DG—obtained by filling in a prompt template with specific parameters (for example, the target and taboo words in `Taboo`)—is called a *dialogue game realisation instance*, or simply an *instance*. We denote instance i of game g as $x_{g,i}$. Each instance defines a *game tree*, which originates from the initial game description and branches out at every turn into all possible actions available at that point. In other words, the tree contains all possible gameplays for that particular instance. If the DG allows verbal actions of unbounded length (i.e., compositional and infinite action spaces), the corresponding game tree will have an infinite number of nodes and edges. A (complete) *trajectory* is a path from the root node to a leaf node. Each player in the game is represented by a *policy* π specifying their action at each decision point. When all players in a game are instantiated with policies, this collectively induces a distribution over trajectories. An *episode* of a given instance can be viewed as a sample from this distribution. The resulting interaction can be recorded as a *transcript* t as follows (this description applies to a two-player game, with straightforward extensions to games involving more players).

Definition 3.1 (Transcript)**.** A *transcript* t represents a trajectory through a game instance tree:

$$t = (S_0, C_1^A, R_1^A, S_0', C_1^B, R_1^B, S_1, \ldots, S_{F-1}, C_F^A, R_F^A, S_{F-1}', C_F^B, R_F^B, S_F)$$

where C_n^P is the context shown to player P at turn n, R_n^P is their response, and S_n is the abstract game state at turn n. The initial C_1^P has a special status, as it contains the game description. A

[3] See Chalamalasetti et al. (2023) for an example of a full prompt that can be used, which needs to contain additional formatting instructions.

trajectory is *complete* if it ends in a final state S_F or an abort state S_X. Dialogue Game Feedback is given by a game-specific feedback function f which evaluates trajectories. The scoring function typically evaluates complete trajectories, although in certain games—including some of the games described below—incomplete trajectories may also be assessed.

A transcript t as defined above contains all interleaved interactions between players and the Game Master. However, each player P only observes the contexts C^P presented to them, never the raw responses of other players. To recover a player's view, a perspective function p_P reduces a trajectory to the sequence of context–response pairs (C_i^P, R_i^P) for $0 \leq i < n$ that player P has experienced at turn n.

3.3 The PLAYPEN Environment

The PLAYPEN environment we introduce here builds upon the Dialogue Game benchmark (`clembench`; Chalamalasetti et al., 2023), transforming it into an interactive playground in which LLMs can learn to be language users. At the time of writing, PLAYPEN includes 15 `clembench` Games, testing language and world knowledge (e.g., in games such as `Taboo`, `Wordle`, `Codenames`), the ability to perform conversational grounding (e.g. `PrivateShared`, `GuessWhat`), and spatial and causal reasoning (e.g., `Adventure Games` or `Map Navigation`). We provide the full list of games with further details in Appendix A.

By recording trajectories as defined in Section 3.2, PLAYPEN supports both offline and online learning, as well as the representation of branching subtrees within the overall game tree through repeatedly sampling from player policies. This flexibility enables the learning experiments that we turn to now.

4 Experimental Setup

Our experiments focus on leveraging Dialogue Game Feedback to post-train language models, building on the premise that this feedback signal is most effectively used when the model is already capable of prompted gameplay to a certain extent.

4.1 Models

We selected Llama-3.1-8B and Llama-3.1-70B (Meta, 2024), both in the `Instruct` variant, as they have generally shown to be receptive to further training (Taori et al., 2023; Zhou et al., 2023a), and have performed well within their size classes on the public `clembench` leaderboard.[4] We also report results for Qwen2-7B (Qwen-2 Team, 2024) (also in its `Instruct` variant) to showcase the effect of the different training regimes on another state-of-the-art LLM as well. In some experiments, we also used a 4-bit quantised version of the model for more efficient training and inference.[5]

4.2 Evaluation

What improvements can we expect from learning with Dialogue Game Feedback? We hypothesise that we will see improvements in gameplay on unseen instances of the games encountered during training, as well as generalisation to new game types. To assess the broader impact of DFG learning, we additionally evaluate the post-trained models for their formal and functional linguistic competence, as well as on general NLP benchmarks.

Interactive Dialogue Games. Performance on interactive dialogue games is evaluated using the `clemscore` metric (Chalamalasetti et al., 2023), which captures both the ability to adhere to the formal rules of a game and the quality of the gameplay. Specifically, the `clemscore` is obtained by multiplying the macro-average percentage of games that were validly played with the macro-average quality score (typically task success) in those valid attempts. We use 7 of the 15 available Dialogue Games for training: `Taboo`, `PrivateShared`, `ImageGame`, `ReferenceGame`, and three variants of `Wordle`. For evaluation, we generate new instances of these games to form the *in-domain* test set. The remaining eight games—`Codenames`, `Adventure Game`, `GuessWhat`, `MatchIt`, and three variants of `Map Navigation`—serve as an additional *out-of-domain* test set.

Formal Linguistic Competence. We evaluate formal linguistic competence (Mahowald et al., 2024), such as the ability to recognise morphosyntactic agreement or lexical entailment, using the GLUE Diagnostic dataset (Wang et al., 2018).

Functional Linguistic Competence. We evaluate cognitive abilities required for verbal interaction, such as working memory, common-sense reasoning, theory of mind and other executive and socio-emotional skills, using a subset of the tasks

[4] https://clembench.github.io/leaderboard.html, accessed 2025-05-09.

[5] https://huggingface.co/unsloth/Meta-Llama-3.1-8B-Instruct-bnb-4bit

curated by Momentè et al. (2025). Specifically, we use a sample from Natural Plan (Zheng et al., 2024), LogiQA 2.0 (Liu et al., 2023), CLadder (Jin et al., 2023), WinoGrande (Sakaguchi et al., 2021), EQ-Bench (Paech, 2023), LM-Pragmatics (Hu et al., 2023), SocialIQA (Sap et al., 2019), and SimpleToM (Gu et al., 2025).

Knowledge and Instruction Following. Finally, we evaluate models on two widely used LLM benchmarks: MMLU-Pro (Wang et al., 2024b) and Big-bench Hard (BBH; Suzgun et al., 2023; Srivastava et al., 2023). In addition, we report performance on IFEval (Zhou et al., 2023b), which assesses general instruction-following capabilities.

4.3 Training Regimes and Interaction Data

We experiment with several training regimes enabled by PLAYPEN, and compare their outcomes with the unmodified Llama-3.1 (`Baseline`). Additional details on training setups, hyperparameters, and data generation procedures are in Appendix E.

4.3.1 Imitation Learning (SFT)

We begin by investigating the potential of pure imitation learning through supervised fine-tuning (SFT). We create a training dataset $\mathcal{D}$ by collecting episodes of gameplay from a large variety of models listed on the public `clembench` leaderboard. For our experiments with SFT, we filter $\mathcal{D}$ to retain only the 7079 successful episodes (i.e., we discard lost and aborted episodes) and convert the transcripts into separate trajectories for all player perspectives: $\mathcal{D}_{SFT} = \{p_A(t), p_B(t) \mid t \in \mathcal{D}, s(t) > \tau_g\}$, for a game-specific threshold τ_g, and using player-specific perspective functions p_A and p_B (see Section 3.2). We fine-tune the `Baseline` model using different data mixtures, containing only interaction data or also instruction-following examples.

After initial experimentation with plain SFT on the 7079 transcripts over 1100 training steps,[6] we found that more sample-efficient variants offered better generalisation. In what follows, we report results from the most effective configuration, which we refer to as `SFT (Cold Start, CS)`. This variant uses only 700 training steps and focuses exclusively on interaction data.

4.3.2 Direct Alignment

We apply Direct Preference Optimization (DPO; Rafailov et al., 2023) to the best SFT checkpoint, `SFT (CS)`; DPO offers a middle ground between SFT and full online reinforcement learning by leveraging contrastive learning on offline data. We consider two variants of DPO training, using dialogue-level or turn-level preference pairs, respectively.

`DPO Dialogue`. For every positive sample in the filtered dataset $\mathcal{D}_{SFT}$, we find a negative counterpart in $\mathcal{D}$ that starts from the same initial state (prompt and game instance S_0) but ends without reaching a successful final state. This includes both aborted and completed but unsuccessful dialogues. The dataset consists of ca. 10K pairs of positive and negative trajectories.

`DPO Turn`. For each turn in a successful dialogue, we find a negative counterpart that shares the same conversational history (the prompt, the game instance and the history up to a branching point) to yield $S_0, C_1^A, R_1^A, \ldots C_n^A, (R_n^A, R_n'^A)$, i.e., paired samples identical up to the penultimate branching node in the game tree. The dataset consists of ca. 86K pairs of positive and negative trajectories.

4.3.3 Online Learning

While both SFT and DPO can provide useful learning signals for gameplay, neither method captures the interactive nature of dialogue games. For this reason, we also performed experiments using GRPO (Shao et al., 2024). Unlike the above methods, GRPO does not rely on a fixed, offline (and possibly off-policy) training dataset $\mathcal{D}$. Instead, for each game instance x_i (initial prompt), we use the target model to interactively produce 8 samples of full gameplay (with temperature set to 0.75). This leads to a group G_i of alternative transcripts. Each trajectory $g \in G_i$ is evaluated using a game-specific reward function that corresponds to the quality score computation for that game in the `clembench` benchmark (see Appendix D.4). We test two configurations of GRPO, with training starting either from the base or the `SFT (CS)` model.

5 Results

We now present the results of running our evaluation suite (Section 4.2) on the post-trained models. Our main finding is that only GRPO learned from Dialogue Game Feedback in a way that generalised to unseen games, without degrading other skills. This is the sole interactive, turn-based method tested. Other methods appeared to overfit to the training distribution and failed to transfer.

[6] Details on this analysis are reported in Appendix E.1.

Model	In Domain			Out of Domain		
	Clemscore =	% Played ×	Quality	Clemscore ↑	% Played ↑	Quality ↑
Llama-3.1-8B						
Baseline	19.39	58.50	33.15	24.58	54.53	45.08
SFT (CS)	40.11	70.48	56.91	22.53	50.55	44.58
SFT (CS) + DPO (Dial.)	32.33	63.54	50.89	19.50	51.67	37.74
SFT (CS) + DPO (Turn.)	39.65	71.28	48.90	23.81	43.03	54.76
GRPO	24.89	57.55	43.25	33.92	67.38	50.34
SFT (CS) + GRPO	24.30	63.22	38.44	31.81	67.26	47.29
Llama-3.1-70B						
Baseline	37.24	64.57	57.67	47.37	82.29	57.57
SFT (CS)	53.60	81.57	65.71	54.40	85.57	63.57
SFT (CS) + DPO (Dial.)	36.92	52.44	70.41	45.46	73.66	61.71
SFT (CS) + DPO (Turn)	38.68	67.59	57.20	50.65	86.29	58.70
Qwen-2-7B						
Baseline	8.14	32.62	37.91	25.15	43.57	36.94
SFT (CS)	32.99	58.56	43.12	19.99	42.21	38.17
SFT (CS) + DPO (Dial.)	10.82	28.58	30.11	13.20	24.08	39.69
SFT (CS) + DPO (Turn.)	22.31	45.21	39.09	21.81	40.65	32.89

Table 1: **Gameplay results.** Clemscore, average percentage of completed games, and average quality score.

5.1 Interactive Gameplay

Table 1 shows dialogue game performance across all evaluated models. The Llama-3.1-8B `Baseline` demonstrates basic rule-following, completing just over half of the games in both in-domain and out-of-domain settings; however, its gameplay quality is generally low. Interestingly, the `Baseline` achieves higher quality and `clemscore` on out-of-domain games, suggesting that the in-domain set may be inherently more challenging for this model. Training with SFT on successful episodes improves in-domain performance but reduces generalisation, with a decrease in all dimensions of performance on out-of-domain games. This aligns with prior observations that SFT tends to overfit to the training distribution (Zeng et al., 2023; Chu et al., 2025; Setlur et al., 2025). Turning to DPO, we observe that both its variants (with dialogue- or turn-level preference pairs) improve in-domain performance over the `Baseline` but fail to outperform the best SFT model. Moreover, they suffer from even stronger degradation in out-of-domain games. We believe this could be a result of "likelihood displacement"—a weakness of DPO-based training strategies (Razin et al., 2024). We applied SFT and DPO to Qwen-2-7B as well. First of all, we notice that the baseline version is less performant than Llama. Despite this difference, we observe a similar trend with respect to the post-training methods, with SFT and DPO showing an improvement over the baseline for in-domain games, but not on out-of-domain games. The most robust training regime is GRPO. Applied directly on the base model, it leads to consistent improvements ($+5.50$ in in-domain `clemscore` and $+9.34$ out-of-domain), albeit with a slight decrease in the number of in-domain games played. When applied on top of SFT, GRPO recovers from this slight drop, likely leveraging the SFT model's stronger ability to adhere to game instructions, though at the cost of more modest gains in quality score.

What happens at the 70B scale? Larger models are known to possess stronger instruction-following capabilities, a skill that is especially relevant for our benchmark, where accurate interpretation of game prompts is critical to gameplay. We therefore conduct additional experiments with `Llama-3.1-70B`, applying the same training regimes used for the 8B model, but excluding GRPO due to its high computational cost and our resource constraints. It is worth noting that simply using this model as a starting point nearly doubles the overall `clemscore` on both in-domain and out-of-domain games (results are shown in the bottom half of Table 1). When applying SFT to the larger base model, we observe diminishing returns for in-domain games, with an improvement of 16.36 points over the `Baseline`—lower than the increase obtained when applying SFT on the base 8B model (20.72). However, on out-of-domain games, we

	In Domain	Out of Domain	Functional, Formal, General, Instruction Following				
Model	**Clemscore**	**Clemscore**	**Executive**	**Socio-Emo**	**GLUE D.**	**General QA**	**IFEval**
Llama-3.1-8B							
Baseline	19.39	24.58	39.24	57.16	38.06	41.86	76.88
SFT (CS)	40.11	22.53	39.93	59.51	40.43	29.95	67.25
SFT (CS) + DPO (Dial.)	32.33	19.50	38.50	55.10	36.20	26.57	68.39
SFT (CS) + DPO (Turn.)	39.65	23.81	39.90	59.75	37.83	31.61	70.00
GRPO	24.89	33.92	39.39	57.51	38.68	41.52	76.67
SFT (CS) + GRPO	24.30	31.81	33.35	58.67	37.31	42.82	75.77
Llama-3.1-70B							
Baseline	37.24	47.37	52.42	71.37	46.16	60.56	85.16
SFT (CS)	53.60	54.40	55.17	69.25	47.72	44.91	79.38
SFT (CS) + DPO (Dial.)	36.92	45.46	48.94	67.89	37.73	38.78	82.26
SFT (CS) + DPO (Turn)	38.68	50.65	50.22	70.21	39.23	44.86	85.68
Qwen-2-7B							
Baseline	08.14	25.04	40.77	56.58	59.68	16.83	59.16
SFT (CS)	32.99	19.99	40.42	56.16	52.98	21.36	53.87
SFT (CS) + DPO (Dial.)	10.82	13.20	39.31	56.79	53.74	15.35	54.20
SFT (CS) + DPO (Turn.)	22.31	21.81	39.22	58.83	48.34	13.71	54.99

Table 2: **Main results**. Clemscore, average percentage of completed games, and average quality score. We report the best SFT variant Cold Start (CS). Executive includes LogiQA 2.0, CLadder, and WinoGrande. Socio-emotional includes EQ-Bench, LM-Pragmatics, SocialIQA, and SimpleToM. General QA includes MMLU-Pro and BBH, while IFEval targets instruction-following specifically. Formal capabilities are evaluated in GLUE Diagnostics. Colored bars indicate whether there is a positive (*green*) or negative difference (*red*) wrt. the Baseline model.

record our best-scoring model with a `clemscore` of 54.4. To calibrate this result, this is still far below the top leaderboard `clemscore` of 70, achieved by `o3-mini-2025-01-31`.[7] Finally, we find that applying DPO on top of SFT reverses some of the gains of SFT alone. Between the two DPO variants, DPO with dialogue-level preference data obtains lower scores on both in-domain and out-of-domain games—another possible case of overfitting after the preliminary SFT phase. The turn-level variant of DPO yields modest improvements over the `Baseline` on both in-domain and out-of-domain games, but it still falls short of the best SFT model.

5.2 Non-Interactive Benchmarks

While our main focus is on performance in dialogue games, we also evaluate models across a broad set of other tasks. This helps identify whether training on dialogue games leads to regressions in general language skills (e.g., formal competence) or, conversely, contributes to improvements in language use (e.g., functional competence). Table 2 summarises the results across these evaluations.

Among the post-training regimes tested, GRPO is the most balanced overall, with lower oscillations—either improvements or regressions—in non-interactive task performance. Tables 11 and 12 in Appendix F.4 give a complete overview of the results. Here, we highlight that training on dialogue games seems to provide a modest improvement on the "Executive" task category for the 70B model trained with `SFT (CS)`, suggesting that learning from dialogue games may enhance a model's ability to integrate and reason over contextual information—an ability Momentè et al. (2025) identified as critical for these tasks.

Another relevant finding from this evaluation is the substantial drop in the ability to follow instructions, as measured by IFEval. This calls for further investigation into instruction-following training regimes that are more robust to interactive settings, allowing models not only to handle single-turn prompts—as is common in current instruction-following regimes—but also to participate effectively in complex, goal-oriented, rule-governed, and multi-turn tasks such as dialogue games.

5.3 Qualitative Discussion

An outcome of our evaluation is that increasing the size of the LLM backbone does correspond to better instruction-following abilities in both traditional benchmarks and on our dialogue games.

[7] Based on `https://clembench.github.io/leaderboard.html`, accessed 2025-05-14.

However, current models still do not show desirable online adaptation skills (also see Appendix G for a detailed error analysis on gameplay abilities). Thanks to language prompts describing a game g, we should be able to derive a game policy π_g on the fly. In some cases, this might not be enough, and therefore, it is possible to use SFT to learn how to play the game by mimicking transcripts. However, because the model does not have the chance to play by itself, it might miss some nuances of the game and overfit on specific rules/formats of the game at hand. On the other hand, thanks to the online training regime of GRPO, it is possible to acquire general-purpose instruction following abilities that allow models to perform better in out-of-domain games as well as retain abilities required for more general-purpose NLP tasks—a result in line with test-time compute analysis for RL algorithms reported in the literature (Chu et al., 2025; Setlur et al., 2025). We report additional dialogue transcripts in Table C in the Appendix.

6 Conclusion

In this paper, we explore to what extent synthetic interaction in what we call Dialogue Games—goal-directed and rule-governed activities driven by verbal actions—can provide a learning signal for LLMs. We created PLAYPEN, an environment that facilitates synthetic data generation of dialogue transcripts that can be used to train LLMs. We provide an extensive evaluation of a variety of state-of-the-art post-training methods such as SFT, DPO, and GRPO, and show how GRPO is a more stable training regime that prevents overfitting to in-domain games and facilitates generalisation to out-of-domain dialogue games. Additionally, we demonstrate that when leveraging dialogue games, it is possible to observe a performance improvement when completing more traditional, non-interactive general instruction-following tasks such as MMLU-Pro.

The framework and the baselines presented here can form the basis for exciting future work, for example investigating novel training regimes based on reinforcement learning to truly leverage the multi-turn nature of dialogue games, or exploring the use of intermediate language feedback that can be acquired as part of the interaction (along the lines of Sumers et al. 2021), and further exploring the potential of the "learning in conversational interaction" paradigm.

Limitations

Our study makes significant strides in demonstrating the potential of dialogue games as a valuable source of feedback signals for training LLMs. The PLAYPEN environment offers a versatile platform for exploring both off- and online learning paradigms, and our comparative analysis of post-training methods, including SFT, DPO, and reinforcement learning with GRPO, provides a strong foundation for future research.

However, the current study has several limitations that warrant further investigation. Firstly, in our Direct Preference Optimization (DPO) experiments, we utilize a seed dataset of successful dialogues from which we derive positive and negative pairs. However, for the turn variant, we assume that all turns within these dialogues are successful. This assumption may not hold true in real-world scenarios, particularly when corrections or clarifications are present within the dialogue (Chiyah-Garcia et al., 2024). Secondly, our work does not explore multi-turn training methods, which could be crucial for more complex dialogue games and real-world applications where it is important to perform credit-assignment across multiple turns (e.g., Zhou et al., 2024b).

Additionally, we did not incorporate reasoning models (e.g., DeepSeek-AI et al., 2025) or chain-of-thought prompting techniques (Wei et al., 2022), which have shown promise in enhancing LLM performance in other tasks. Furthermore, our evaluation of GRPO is limited to a smaller 8B LLM. Evaluating the effectiveness of GRPO on larger models, such as the 70B parameter model, would provide valuable insights into the scalability of our findings. Unfortunately, due to limited computational resources, we leave this exploration for future work.

The current set of dialogue games in PLAYPEN provides a foundation for our research, but it is not exhaustive. Future work should aim to expand the set of games to be more representative of the diverse range of language games encountered in real-world scenarios. This is especially important considering that Momentè et al. (2025) has demonstrated that dialogue games are actually more discriminative than other benchmarks because they likely require important underlying capabilities such as working memory.

Finally, our focus is on dialogue game feedback, which is inherent to the task itself. We do not

consider additional forms of feedback, such as explicit corrective feedback, which could potentially enhance learning, as explored in prior work (e.g., Sumers et al., 2021; McCallum et al., 2023; Xi et al., 2024).

Ethical considerations

Our work broadly falls under the rubric of "self-improvement" of language models. There is a small, but non-zero chance that such self-improvement, if run unsupervised and in recursive loops, might lead to uncontrolled gains. Our advice hence would be to define clear stopping criteria for learning runs. Additionally, we created PLAYPEN as a synthetic and simulated learning environment where the model doesn't have access to external tools or, more broadly, it doesn't have the ability to execute actions in the real world.

Acknowledgments

This research has been partially supported and sponsored by the Province of Bolzano and the EU through the project Artificial Intelligence Laboratory (AI-Lab) ERDF 2021-2027 EFRE1047, CUP: I53C23001670009 (01/01/2024 - 31/12/2026). Alessandro Suglia, Raffaella Bernardi, and David Schlangen acknowledge ISCRA for awarding this project access to the LEONARDO supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CINECA (Italy). Mario Giulianelli was in part supported by an ETH Zurich Postdoctoral Fellowship. Alberto Testoni is funded by the project CaRe-NLP with file number NGF.1607.22.014 of the research programme AiNed Fellowship Grants, which is (partly) financed by the Dutch Research Council (NWO). Raquel Fernández was supported by the European Research Council (ERC Consolidator Grant DREAM 819455). David Schlangen acknowledges support by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – grant 423217434 (RECOLAGE), and University of Potsdam (President's Fund).

References

Afra Feyza Akyurek, Ekin Akyurek, Ashwin Kalyan, Peter Clark, Derry Tanti Wijaya, and Niket Tandon. 2023. RL4F: Generating natural language feedback with reinforcement learning for repairing model outputs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7716–7733, Toronto, Canada. Association for Computational Linguistics.

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.

Leonardo Bertolazzi, Davide Mazzaccara, Filippo Merlo, and Raffaella Bernardi. 2023. ChatGPT's information seeking strategy: Insights from the 20-questions game. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 153–162, Prague, Czechia. Association for Computational Linguistics.

Maciej Besta, Julia Barth, Eric Schreiber, Ales Kubicek, Afonso Catarino, Robert Gerstenberger, Piotr Nyczyk, Patrick Iff, Yueling Li, Sam Houliston, Tomasz Sternal, Marcin Copik, Grzegorz Kwaśniewski, Jürgen Müller, Łukasz Flis, Hannes Eberhard, Hubert Niewiadomski, and Torsten Hoefler. 2025. Reasoning Language Models: A Blueprint. *arXiv preprint*. ArXiv:2501.11223 [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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jerome S. Bruner and Rita Watson. 1983. *Child's talk: learning to use language*, 1st ed edition. W.W. Norton, New York. Format: E-Book.

Kris Cao, Angeliki Lazaridou, Marc Lanctot, Joel Z Leibo, Karl Tuyls, and Stephen Clark. 2018. Emergent Communication through Negotiation. In *ICLR 2018*, pages 1–15. ArXiv: 1804.03980.

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.

Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Hu, Jaap Jumelet, Tal Linzen, Jing Liu, Aaron Mueller, Candace Ross, Raj Sanjay Shah, Alex Warstadt, Ethan Wilcox, and Adina Williams. 2025. BabyLM Turns 3: Call for papers for the 2025 BabyLM workshop. *arXiv preprint*. ArXiv:2502.10645 [cs].

Zizhao Chen, Mustafa Omer Gul, Yiwei Chen, Gloria Geng, Anne Wu, and Yoav Artzi. 2024. Retrospective learning from interactions. *arXiv preprint arXiv:2410.13852*.

Javier Chiyah-Garcia, Alessandro Suglia, and Arash Eshghi. 2024. Repairs in a block world: A new benchmark for handling user corrections with multi-modal language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11523–11542.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*.

Eve V. Clark. 2016. *First language acquisition*, 3rd ed edition. Cambridge University Press, Cambridge.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. *arXiv preprint*. ArXiv:2110.14168 [cs].

Marc-Alexandre Côté, Akos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, and 1 others. 2019. Textworld: A learning environment for text-based games. In *Computer Games: 7th Workshop, CGW 2018, Held in Conjunction with the 27th International Conference on Artificial Intelligence, IJCAI 2018, Stockholm, Sweden, July 13, 2018, Revised Selected Papers 7*, pages 41–75. Springer.

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.

Christopher Zhang Cui, Xingdi Yuan, Zhang Xiao, Prithviraj Ammanabrolu, and Marc-Alexandre Côté. 2025. Tales: Text adventure learning environment suite. *arXiv preprint arXiv:2504.14128*.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, and et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *Preprint*, arXiv:2501.12948.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. *arXiv preprint arXiv:2305.14314*.

Jinhao Duan, Renming Zhang, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Elias Stengel-Eskin, Mohit Bansal, Tianlong Chen, and Kaidi Xu. 2024. GTBench: Uncovering the Strategic Reasoning Limitations of LLMs via Game-Theoretic Evaluations. *Preprint*, arXiv:2402.12348.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Raquel Fernández, Staffan Larsson, Robin Cooper, Jonathan Ginzburg, and David Schlangen. 2011. Reciprocal learning via dialogue interaction: Challenges and prospects. In *Proceedings of the IJCAI 2011 Workshop on Agents Learning Interactively from Human Teachers (ALIHT)*, Barcelona, Catalonia.

Zhaolin Gao, Wenhao Zhan, Jonathan D Chang, Gokul Swamy, Kianté Brantley, Jason D Lee, and Wen Sun. 2024. Regressing the relative future: Efficient policy optimization for multi-turn rlhf. *arXiv preprint arXiv:2410.04612*.

Ran Gong, Qiuyuan Huang, Xiaojian Ma, Hoi Vo, Zane Durante, Yusuke Noda, Zilong Zheng, Song-Chun Zhu, Demetri Terzopoulos, Li Fei-Fei, and Jianfeng Gao. 2023. Mindagent: Emergent gaming interaction. *Preprint*, arXiv:2309.09971.

Yuling Gu, Oyvind Tafjord, Hyunwoo Kim, Jared Moore, Ronan Le Bras, Peter Clark, and Yejin Choi. 2025. SimpleToM: Exposing the Gap between Explicit ToM Inference and Implicit ToM Application in LLMs.

Leon Guertler, Bobby Cheng, Simon Yu, Bo Liu, Leshem Choshen, and Cheston Tan. 2025. TextArena. *arXiv preprint*. ArXiv:2504.11442 [cs].

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.

Kshitij Gupta, Benjamin Thérien, Adam Ibrahim, Mats L. Richter, Quentin Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. 2023. Continual pre-training of large language models: How to (re)warm your model? *CoRR*, abs/2308.04014.

Betty Hart and Todd R. Risley. 1995. *Meaningful differences in the everyday experience of young American children*. Paul H. Brookes Publishing Co.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring Mathematical Problem Solving With the MATH Dataset.

Sarah Hiller and Raquel Fernández. 2016. A data-driven investigation of corrective feedback on subject omission errors in first language acquisition. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 105–114, Berlin, Germany. Association for Computational Linguistics.

Arian Hosseini, Xingdi Yuan, Nikolay Malkin, Aaron Courville, Alessandro Sordoni, and Rishabh Agarwal. 2024. V-STaR: Training Verifiers for Self-Taught Reasoners.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. 2023. CLadder: Assessing causal reasoning in language models. In *NeurIPS*.

Patricia K. Kuhl. 2007. Is speech learning 'gated' by the social brain? *Developmental Science*, 10(1):110–120.

Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip H. S. Torr, Salman Khan, and Fahad Shahbaz Khan. 2025. LLM Post-Training: A Deep Dive into Reasoning Large Language Models. *arXiv preprint*. ArXiv:2502.21321 [cs].

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Nathan Lambert. 2024. *Reinforcement Learning from Human Feedback*. Online.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, and 4 others. 2025. Tülu 3: Pushing Frontiers in Open Language Model Post-Training. *arXiv preprint*. ArXiv:2411.15124 [cs].

Jiatong Li, Rui Li, and Qi Liu. 2023. Beyond Static Datasets: A Deep Interaction Approach to LLM Evaluation. *Preprint*, arXiv:2309.04369.

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.

Hanmeng Liu, Jian Liu, Leyang Cui, Zhiyang Teng, Nan Duan, Ming Zhou, and Yue Zhang. 2023. Logiqa 2.0—an improved dataset for logical reasoning in natural language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2947–2962.

Ziqiao Ma, Zekun Wang, and Joyce Chai. 2025. Babysit A Language Model From Scratch: Interactive Language Learning by Trials and Demonstrations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 991–1010, Albuquerque, New Mexico. Association for Computational Linguistics.

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*.

Davide Mazzaccara, Alberto Testoni, and Raffaella Bernardi. 2024. Learning to ask informative questions: Enhancing LLMs with preference optimization and expected information gain. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5064–5074, Miami, Florida, USA. Association for Computational Linguistics.

Sabrina McCallum, Max Taylor-Davies, Stefano Albrecht, and Alessandro Suglia. 2023. Is feedback all you need? leveraging natural language feedback in goal-conditioned rl. In *NeurIPS 2023 Workshop on Goal-Conditioned Reinforcement Learning*.

Meta. 2024. The Llama 3 Herd of Models.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. To appear.

Jens Nevens and Michael Spranger. 2017. Computational models of tutor feedback in language acquisition. In *2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pages 221–226. IEEE.

Mitja Nikolaus and Abdellah Fourtassi. 2021. Modeling the interaction between perception-based and production-based learning in children's early acquisition of semantic knowledge. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 391–407, Online. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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. *Preprint*, arXiv:2308.10032.

Qwen-2 Team. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Noam Razin, Sadhika Malladi, Adithya Bhaskar, Danqi Chen, Sanjeev Arora, and Boris Hanin. 2024. Unintentional unalignment: Likelihood displacement in direct preference optimization. *arXiv preprint arXiv:2410.08847*.

Philipp Sadler, Sherzod Hakimov, and David Schlangen. 2024. Sharing the cost of success: A game for evaluating and learning collaborative multi-agent instruction giving and following policies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14770–14783, Torino, Italia. ELRA and ICCL.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Matthew Saxton. 2000. Negative evidence and negative feedback: Immediate effects on the grammaticality of child speech. *First Language*, 20(60):221–252.

David Schlangen. 2019. Language tasks and language games: On methodology in current natural language processing research. *CoRR*, abs/1908.10747.

David Schlangen. 2023. Dialogue games for benchmarking language understanding: Motivation, taxonomy, strategy. *CoRR*, abs/2304.07007.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. 2024. Rewarding Progress: Scaling Automated Process Verifiers for LLM Reasoning.

Amrith Setlur, Nived Rajaraman, Sergey Levine, and Aviral Kumar. 2025. Scaling Test-Time Compute Without Verification or RL is Suboptimal. *arXiv preprint*. ArXiv:2502.12118 [cs].

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Chandler Smith, Rakshit Trivedi, Jesse Clifton, Lewis Hammond, Akbir Khan, Sasha Vezhnevets, John P. Agapiou, Edgar A. Duéñez-Guzmán, Jayd Matyas, Danny Karmon, Marwa Abdulhai, Dylan Hadfield-Menell, Natasha Jaques, Joel Z. Leibo, Oliver Slumbers, Tim Baarslag, and Minsuk Chang. 2024. The Concordia Contest: Advancing the Cooperative Intelligence of Language Agents.

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 Tazarv, and 431 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*. Featured Certification.

Luc Steels. 2003. Evolving grounded communication for robots. *Trends in Cognitive Sciences*, 7(7):308–312.

Alessandro Suglia, Ioannis Konstas, 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.

Theodore R Sumers, Mark K Ho, Robert D Hawkins, Karthik Narasimhan, and Thomas L Griffiths. 2021. Learning rewards from linguistic feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6002–6010.

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.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. 2024a. MINT: Evaluating LLMs in Multi-turn Interaction with Tools and Language Feedback. In *The Twelfth International Conference on Learning Representations*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned Language Models Are Zero-Shot Learners. pages 1–46. ArXiv: 2109.01652.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 24824–24837.

Yue Wu, Xuan Tang, Tom M. Mitchell, and Yuanzhi Li. 2024. SmartPlay: A Benchmark for LLMs as Intelligent Agents. *Preprint*, arXiv:2310.01557.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-Grained Human Feedback Gives Better Rewards for Language Model Training.

Jiajun Xi, Yinong He, Jianing Yang, Yinpei Dai, and Joyce Chai. 2024. Teaching embodied reinforcement learning agents: Informativeness and diversity of language use. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4097–4114.

Aohan Zeng, Mingdao Liu, Rui Lu, Bowen Wang, Xiao Liu, Yuxiao Dong, and Jie Tang. 2023. Agenttuning: Enabling generalized agent abilities for LLMs. *arXiv preprint arXiv:2310.12823*.

Huaixiu Steven Zheng, Swaroop Mishra, Hugh Zhang, Xinyun Chen, Minmin Chen, Azade Nova, Le Hou, Heng-Tze Cheng, Quoc V. Le, Ed H. Chi, and Denny Zhou. 2024. NATURAL PLAN: Benchmarking LLMs on Natural Language Planning.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, and 1 others. 2023a. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023b. Instruction-following evaluation for large language models. *Preprint*, arXiv:2311.07911.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024a. SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents. In *Proceedings of ICLR 2024*, pages 1–45.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. 2024b. ArCHer: Training Language Model Agents via Hierarchical Multi-Turn RL. *arXiv preprint*. ArXiv:2402.19446 [cs].

Appendix

In the following, we report additional details for our paper, starting from the dialogue games used in our work (see Section A for all the available games and Section C for some examples). We also provide details of our experimental protocol, including training data (see Section D), our training regime (see Section E), and additional results (see Section F). Beyond raw scores on the different benchmarks that are part of Playpen, we also provide a comprehensive error analysis to understand the behaviour of each model after our post-training regime via Playpen (see Section G). Finally, we also report licenses for our code and data (see Section B).

A Dialogue Games

- **Taboo**: two-player game where one player gives a clue, not using certain words, and the other player needs to guess a target word based on the clue.
- **PrivateShared**: two-player scorekeeping game where an answerer agent goes through a form with a questioner. The GameMaster

keeps track of which information has been already shared and checks whether players also do so correctly.

- **ImageGame**: two-player instruction giving and following game where one player provides a description of a target image (a matrix in ASCII format) and the other player has to draw (i.e. generate some matrix again in ASCII format) an image based on the description. The generated image should match the target one.
- **ReferenceGame**: two-player game where one player is given three different images (one of them is selected as the target) and has to generate a referring expression that describes the target image by differentiating it from other two (distractors). Another player is then given the same images (orders are shuffled) and is asked to guess which one is the target based on the given referring expression. It is the only single-turn game in the benchmark.
- **Wordle**: popular single-player game where the task is to guess a 5-letter word. In each turn, feedback is provided based on the placements of characters in the guessing attempts.
- **Wordle With Clue**: slightly changed version of the base Wordle game with the addition of a clue for the target word.
- **Wordle With Critic**: a two-player version of the base Wordle game where the second player (critic) provides feedback on the guesses of the first player.
- **Codenames**: a popular cooperative game with two teams that try to uncover their agent's code names (words). Here, one player has to describes clues that could strategically correspond to more than one word. Teams are composed of a Spymaster that provides clues and a Field operative tasked with guessing. In our case, two LLMs are placed on the same team, with the other team being characterized by programmatic behaviour.
- **AdventureGame**: single-player text adventure game where a player is placed in a random location of an environment and is given a task (e.g. pick up the flower and place it on the table). The player explores the environment by giving commands (e.g. "go to the kitchen", "open the cupboard") and the environment provides feedback regarding whether the command can be realised and its outcome or not. The player explores the world (with multiple rooms and objects in them) and has to decide on its own when to stop the game.
- **GuessWhat**: two-player information seeking game where one player needs to guess the target word (out of eight options) by asking questions about the target. The other player knows the target word and answers the questions with "yes" or "no".
- **Matchit**: two-player game where each player is given an image (an ASCII representation, e.g. grid), which is not revealed to the other player. The goal is for players to understand whether they are looking at the same image or not. Players are allowed to ask questions to each other about the image and provide answers to the questions.
- **Map Navigation**: single-player game where a player has to navigate a map by indicating the directions where to go. Initially, the player is placed in a random room and is given information regarding where adjacent rooms are located (east, west, north, south) and about the name of the current room (e.g kitchen, bathroom, garage). Every time the player makes a step, is provided with these information updated at the current location. When the player thinks it has explored everything in the map, it should decided to stop the navigation on its own.
- **Map Navigation - specific room**: modified version of the game above. The player is given the name of a target room (e.g. kitchen, garage, living room) and is required to stop exploring once it finds the target.
- **Map Navigation** with graph reasoning: modified version of the base Mapnavigation game where, along with the already described exploration mechanism, at each step the player is required to draw (i.e. generate JSON-like structure) the graph representing the map discovered up to that point.

B License for Artifacts

PLAYPEN's implementation reuses important components from clemcore, and all the games used

come from clembench (Chalamalasetti et al., 2023). Therefore, we inherit its MIT License for PLAYPEN. We acknowledge that the transcripts used for SFT and DPO training are derived from the clembench-runs [8] repo which has CC-By 4.0 License.

C Dialogue Games Transcripts

We include below sample transcripts for three of the games present in this work: GuessWhat, Taboo, and Codenames. The examples show successful episodes for the Taboo and GuessWhat games, while for Codenames the episode is aborted before reaching the end.

C.1 GuessWhat game Transcript

#	A	Game Master	B
1		[A⟨GM] Let's play a guessing game. Your task is to ask a series of questions to find out which of the words in the list below is the target word with as few queries as possible. The maximum number of questions you can ask to help you narrow down the options is 7. Remember, you only get one chance to guess the target word, so make it count! The rules of the game are: 1. Only ask one question per turn. DO NOT ASK MORE THAN ONE QUESTION. 2. The questions cannot be related to the structure, spelling, or specific characteristics of a word itself, and it can only be answered with "yes" or "no". 3. Use the following format for asking questions and DO NOT ADD ANY TEXT OTHER THAN THIS: QUESTION: <your question> 4. You have only one opportunity to guess the target word. To make your guess, use the following format and DO NOT ADD ANY TEXT OTHER THAN THIS: GUESS: <target word> Remember you cannot ask for a hint. Make your question brief. Let's start! List of candidate words: ['Mackerel', 'Grouper', 'Carp', 'Catfish', 'Halibut', 'Flounder', 'Minnow', 'Goldfish']		
2		[GM⟩B] Let's play a guessing game! Your task is to answer the other player's questions. Based on your knowledge of the word: Catfish, respond to the following questions or guesses. Limit your response to only 'yes' or 'no' with no explanation or other words. Never reveal the answer in your response.	[A⟩GM] QUESTION: Is the target word a freshwater fish?
3		[GM⟩B] You must reply using the format below and DO NOT ADD ANY TEXT OTHER THAN THIS: ANSWER: <some text> Target Word: Catfish QUESTION: Is the target word a freshwater fish?	
4		[GM⟨B] ANSWER: Yes	
5			[A⟩GM] QUESTION: Is the target word commonly kept as a pet?
6		[GM⟨B] ANSWER: Yes	
7			[A⟩GM] QUESTION: Is the target word relatively small in size?
8		[GM⟨B] ANSWER: No	
9			[A⟩GM] QUESTION: Is the target word native to North America?
10		[GM⟨B] ANSWER: Yes	
11			[A⟩GM] QUESTION: Is the target word a scaleless fish?
12		[GM⟨B] ANSWER: Yes	
13			[A⟩GM] QUESTION: Does the target word have whisker-like barbels?
14		[GM⟨B] ANSWER: Yes	
15			[A⟩GM] GUESS: Catfish
16		[GM|GM] CORRECT GUESS	
17		[GM|GM] End game.	

[8] https://github.com/clembench/clembench-runs/blob/main/LICENSE

C.2 Taboo game Transcript

#	A	Game Master	B
1		[A⟨GM] You are playing a collaborative word guessing game in which you have to describe a target word for another player to guess. Rules: (a) You have to reply in the form: CLUE: <some text>. Guesses from the other player will start with GUESS. (b) You cannot use the target word itself, parts or morphological variants of it in your description. (c) In addition, the same rules apply for related words which are provided below. End conditions: (i) If you use the target word or a related word in your description, then you lose. (ii) If the other player can guess the target word in 3 tries, you both win. Let us start. This is the target word that you need to describe and that the other player needs to guess: ordinary Related words are: - common - normal - plain Important: You are under time pressure, give short descriptions that are to the point!	
2	[A⟩GM] CLUE: Something that is usual or expected.		
3		[GM⟩B] You are playing a collaborative word guessing game in which you have to guess a target word that another player describes to you. You can make one guess at each trial. You win when you guess the target word. You lose when you cannot guess it in 3 tries. After each trial you will get a new hint from the other player which starts with CLUE. Make your guesses by just saying the word using the following form: GUESS: <a word> Let us start. CLUE: Something that is usual or expected.	
4		[GM⟨B] GUESS: Norm	
5	[A⟩GM] CLUE: Not fancy or special.		
6		[GM⟨B] GUESS: Ordinary	
7		[GM|GM] CORRECT GUESS	
8		[GM|GM] End game.	

C.3 Codenames game Transcript

#	A	Game Master	B
1		[A⟨GM] Provide one single-word clue that relates to some of your team's words from the list below. You can choose to how many and to which words your clue relates to. The goal is to have your partner guess all of your team's words based on the clues you give them, before the other team has uncovered all of theirs. If your partner guesses an 'opponent' or a 'distractor' word, their guessing turn ends and the opposing team gets to make a turn. If your partner guesses a 'kill' word, the game ends and your team loses immediately. The more words your clue relates to, the quicker you can win the game, but the harder the associations may be to guess for your partner, so choose your clue word wisely. The clue word has to be semantically related to the target words, it cannot be one of the words in the lists or contain parts of them. Always give your single-word clue and your comma-separated list of related target words in the following format and make your answers as short as possible, never include any other text than is required in this form: CLUE: <WORD> TARGETS: <TARGETS> Your team words are: hail, neck, mountain, poland, sister, couch, stool, captain, officer. Your opponent's team words are: hair, teeth, england, sergeant, table, fog, desert, rye. Distractor words are: daughter, recliner, private, son, cousin, thigh, ireland. The kill words are: liquor.	

2 [A⟩GM] CLUE: storm
TARGETS: hail, mountain

[GM⟩B] Provide a comma-separated list of up to 2 words from the following list that best relate or are most closely associated with the word '**storm**'.
Always start your list of guess(es) with 'GUESS: ' and do not include any other text
3 in your answer.

sister, hair, hail, couch, table, desert, recliner, poland, daughter, thigh, teeth, cousin, stool, sergeant, private, england, son, fog, ireland, liquor, neck, officer, rye, mountain, captain

4 [GM⟨B] GUESS: hail, storm

[GM|GM] Guessed word **storm** is the
5 same word as the provided clue word (storm), you should only select words from the provided list

6 [GM|GM] Game aborted.

D Training Data Generation

D.1 Supervised Fine-tuning

Data Collection We preprocessed the dialogue transcripts available in the clembench-runs repository[9] and transformed them into a format that allowed further filtering. The clembench-runs repository is a collection of all generated interaction data produced by the models evaluated on the benchmark. The extraction included the removal of all GameMaster-specific logging and parsing prompts that were not part of the conversation pieces directly prompted to the model. Besides the conversational data, we also extracted additional metadata such as the information about the model, played game, benchmark-version, experiment, and episode number, as well as whether the episode was successfully played, lost or aborted.

Data Filtering For the supervised fine-tuning, only successful episodes were considered for training, with lost and aborted ones discarded from the data.

Data Transformation After filtering, we added some game-specific data transformations to mitigate changes between the different benchmark versions and to improve training performance. Most of the transformations were necessary due to changes in the prompts between the benchmark versions and changes in the parsing rules for model answers. All of the transformations and associated justification are listed below:

ImageGame While clembench versions 0.9 and 1.0 allowed the player to add *"what is your next instruction"* to its answers, the same behavior led to parsing errors in the version 1.6, which resulted in the abortion of all ImageGame episodes. To address this problem, all model answers from player-1 were truncated to only contain the correct format required by the version 1.6.

```
Before: Instruction: Put a B in the
first column of all rows
what is your next instruction

After: Instruction: Put a B in the
first column of all rows
```

Wordle and its variants For the three Wordle variants, there were a few successfully played episodes that contained an *"INVALID_FORMAT"* token inside the prompts. These instances were removed since the model should not reproduce outputs with invalid formats. We have also removed the parts of episodes where the player is asked to provide again an answer after the providing an unparsable one (Wordle-specific).

ReferenceGame For ReferenceGame, the initial prompt was changed from the older clembench versions (0.9 and 1.0) to version 1.6. While the older versions contained multiple examples (few-shot prompting), in version 1.6 there are no examples available. These examples directly biased the model towards a specific gameplay strategy, and provide a description of the grids (see following snippet from the old version of the prompt).

```
Here is an example with grids.
The first grid is the target grid and
the following two grids are distractors.

Target grid:
X X X X X
O O X O O
O O X O O
O O X O O
O O X O O
...

The referring expression for the given
target grid is like so:
Expression: Filled as T.
```

Here, the model is directly instructed to describe the whole grid as a letter or shape. However, not all grids follow this pattern. In addition, considering that the game is a two-player reference game, player-2 has a 33% chance of guessing correctly the correct grid. This resulted in a situation where about 53% of the successful episodes, player-1 described the target grid as "Filled as T" while, except for the prompt example, there is no T-shaped grid in the data. This meant that reference game data from the old benchmark versions could not be used

[9] https://github.com/clembench/clembench-runs

for the training process due to the low quality. To mitigate this problem, data from the version 1.6 was used while 20-30% of the episodes of each experiment were held out for testing.

PrivateShared For PrivateShared, after the first experiments it appeared that in most cases, the trained model answered with *"ASIDE: No"* to all probe-questions. In PrivateShared, the model should act like a customer at a travel agency that wants to travel. The agent asks questions about destination, time and other related properties of the inquiry. Turn after turn, the model has to tell the agent all the information the agent needs. After every question there is a block of probing questions where the model is asked whether or not specific information has been shared already, and the model has to answer with *"ASIDE: yes"* or *"ASIDE: no"*.

Considering the structure of the game, the model has to answer with *"ASIDE: no"* to all probe-questions in the beginning of the game dialogue which changes to more and more *"ASIDE: yes"* during the course of the game play depending on what information has already been shared. To prevent over-fitting, we have decided to reduce the number of probe-question answer pairs shown to the model by merging them into a single sample per iteration done in the main conversation.

The specific changes made to PrivateShared, ImageGame and ReferenceGame were partially derived by experimenting with them. For the rest of the games, no particular changes have been made.

Iterative Data-Processing While some of the previously described data transformations were motivated by observations made during the data preparation and collection phase, further experiments have been conducted to iteratively improve the data to optimize fine-tuning performance. This resulted in a final dataset which combined all the possible improvements learned through this process.

In total, more than 30 different experiments were conducted with different dataset configurations. They have been structured into nine main experiments, composed of one or more sub-experiments. Below is a description of the main experiments:

D1 Contains, as an initial experiment, all successfully played episodes of all models. The dialogues are not processed in any way and just parsed into the model-specific chat-templates.

D2 Contains only successfully played episodes from the top k models. The tier list was derived from having the most successful episodes. This directly reflects the models with the best clemscores.

It is to be expected that the quality of the played episodes from better models is higher than the models that only succeeded in a small number of episodes. The idea behind this experiment was to determine whether the difference in quality is reflected by the fine-tuned model.

It appears that training only on the successfully played episodes of the top 10 models has a positive impact on the quality score compared to using all available data.

D3 In the previous experiments, a training sample consisted of a complete episode. This means, the whole conversation over multiple turns was considered as one sample. This implies that intermediate turns were not available as individual training samples in the data. This experiment was designed to determine the impact of using individual conversation pieces as training samples rather than the whole conversation at once. Therefore, every episode was split into individual continuously growing training samples that started with the first question answer pair, and then extended with each question answer pair until the end of the conversation was reached. It is important to note that the data was shuffled before splitting to ensure that conversation parts of one episode remained in the correct order and are trained on together.

We have observed that most of the experiments from D3 outperformed the respective ones from D1 and D2.

D4 was conducted to test different balancing strategies. In the previous experiments, the dataset was not balanced between games. Data can be balanced before or after splitting the conversation parts (as described in D3). The downsampling can be done by random selection or by considering the models' leaderboard positions. Furthermore, there can be oversampling for games with only few available episodes. While balancing overall showed a positive impact, the best performance were achieved when the data was balanced before splitting. The sampling was based on the leaderboard without oversampling. This was also demonstrated by D2, where using the data from the best-performing models showed a positive impact on the fine-tuned models' performance.

D5 & D6 These were two complementary experiments where for D5 the model is only trained on one game, while D6 consists of the opposite experiment and can be described as leave-one-game-out. While this experiment did not yield meaningful in-

sights into the dataset configuration, it led to some improvements with respect to overfitting of the probe questions in PrivateShared.

D7 & D8 were defined to verify or reject the possible improvements derived from game-specific data-transformations. This includes the transformations on PrivateShared and ReferenceGame.

D9 While D1-D8 were completely focused on the data, D9 comprises a hyperparameter tuning of the QLoRA parameters.

Game	Samples Train	Samples Test
Before Splitting		
Taboo	434	18
Referencegame	324	36
Wordle	230	5
Wordle With Critic	302	12
Wordle With Clue	295	5
ImageGame	278	12
PrivateShared	214	5
After Splitting		
Taboo	560	22
ReferenceGame	324	36
Wordle	1038	19
Wordle With Critic	1,192	105
Wordle With Clue	717	12
ImageGame	1,579	52
PrivateShared	1,669	45
Total	7,079	291

Table 3: Final Dataset composition Before and After Splitting.

Final Dataset Overview: As shown in Table 3, depending on the game, the number of samples after splitting varies heavily. Eventually, the total number of samples available for the training is about 7000, while the number of samples for evaluation during the training is about 300.

D.2 Synthetic datasets for warm start and rehearsal training regimes

In order to reinforce and improve the instruction-following capabilities of models during fine-tuning, we designed a synthetic dataset to use for training regimes such as warm-up training and rehearsal training. In contrast to the data derived from clembench runs, this data consists of single-turn user-assistant interactions, and it was programmatically created with the help of human-made templates. These templates consist of short representations of instruction-answer-interactions which we we call 'minigames'. An example of a minigame is the following:

```
[{"role": "user", "content": "Sum these
numbers: 14, 26, and give the answer
after  the tag SUM:"}, {"role":
"assistant", "content": "SUM: 40"}]
```

We included several minigames based on letters/words, numbers, requiring the transformation of inputs into JSON format, or requiring to make choices between different options. In minigames, the focus was on form rather than content given that we aimed through them to enhance models' capabilities of following game instructions regarding the formatting of answers or the shape of input strings. The final dataset consisted of around 20000 samples, obtained by filling the slots of 26 templates. The models' answers to the minigames are required to be short and straight to the point. Only in one case, the minigame is shaped as a multi-turn task. Here, at each turn the model has to select an item from a list only if it has not been chosen already during previous turns. The goal here was that of to enhancing attention to the overall context. The full minigame dataset is released with this work[10].

D.3 DPO

DPO requires paired preference data, i.e., samples sharing the same context before positive vs negative continuations. For DPO Dialogue and DPO Turn with clembench runs, positive continuations are obtained from successful games' interactions, and negative continuations from unsuccessful and aborted games' interactions. Since the SFT models obtained top performances in % Played for all games except Wordle and its variants, we only integrated aborted interactions for this game (and variants). Unsuccessful and aborted interactions have been collected and transformed from the same sources and with the same procedure as SFT data. For multi-player games like Taboo, data for both player 1 (i.e., giving clues) and player 2 (i.e., making guesses) have been integrated into training. The DPO Dialogue dataset consists of around 10K samples as in Fig. 3; the DPO Turn dataset consists of around 58K samples as in Fig. 4.

For DPO Dialogue, we experiment with two variables: the number of negative samples per positive sample and the model source for negative samples. Each positive interaction is paired with n unsuccessful – and n aborted interactions for Wordle and variants – where n is manipulated to find the optimal number of negative trajectories to learn from.

[10] https://huggingface.co/datasets/clembench-playpen/warm-up_synthetic-data.

Prompt

You are playing a collaborative word guessing game in which you have to describe a target word for another player to guess. Rules:
(a) You have to reply in the form: CLUE: <some text>. Guesses from the other player will start with GUESS.
(b) You cannot use the target word itself, parts or morphological variants of it in your description.
(c) In addition, the same rules apply for related words which are provided below.
End conditions:
(i) If you use the target word or a related word in your description, then you lose.
(ii) If the other player can guess the target word in 3 tries, you both win.
Let us start. This is the target word that you need to describe and that the other player needs to guess: transit
Related words are:
- transport
- cross
- traverse

Important: You are under time pressure, give short descriptions that are to the point!

Chosen:

Turn 1:
CLUE: Movement of people or things from one place to another.
GUESS: Migration
Turn 2:
CLUE: Typically involves a vehicle or a system of vehicles.
GUESS: Transportation
Turn 3:
CLUE: Often has fixed routes and schedules.

Rejected:

Turn 1:
CLUE: Movement of people or goods from one place to another, often using public services.
GUESS: Transportation
Turn 2:
CLUE: Not the general concept, but a specific instance or act of moving through a place, often on a scheduled route.
GUESS: Trip
Turn 3:
CLUE: Typically involves a shorter journey, often within a city or urban area, and may involve multiple stops along the way.

Figure 3: **DPO dialogue** dataset: the initial state (prompt and game instance) is shared, the chosen and rejected continuations are the remaining turns from the successful and unsuccessful episodes.

Prompt

You are playing a collaborative word guessing game in which you have to describe a target word for another player to guess. Rules:
(a) You have to reply in the form: CLUE: <some text>. Guesses from the other player will start with GUESS.
(b) You cannot use the target word itself, parts or morphological variants of it in your description.
(c) In addition, the same rules apply for related words which are provided below.
End conditions:
(i) If you use the target word or a related word in your description, then you lose.
(ii) If the other player can guess the target word in 3 tries, you both win.
Let us start. This is the target word that you need to describe and that the other player needs to guess: independently
Related words are:
- separately
- alone
- individually

Important: You are under time pressure, give short descriptions that are to the point!
Turn 1:
CLUE: CLUE: Movement from one place to another.
GUESS: Travel

Chosen:

Turn 2:
CLUE: Often involves public systems like buses or trains.

Rejected:

Turn 2:
CLUE: Not exactly, think of it as moving through a system or route.

Figure 4: **DPO turn** dataset: the initial state and conversational history are shared, the chosen and rejected continuations are single turns from the successful and unsuccessful episodes.

The source for negative samples falls into three conditions: negative samples from all the models, from only the 10 best-performing models, and only models from the same family as the model to train. The best performances, as tested on clembench version *1.6*, have been obtained by coupling 1 negative per positive from the 10 best-performing models. Since the first guess in wordle and wordle_withclue is random, for DPO Turn we restrict to 2k samples from the first turn.

Game	DPO Dialogue	DPO Turn
taboo	4.689	20.074
referencegame	1.712	2.056
wordle_withcritic	1.469	18.234
wordle_withclue	677	3.380
imagegame	1.291	12.094
wordle	285	3.021

Table 4: DPO datasets' samples per game.

D.4 GRPO

Reward Functions. We employed various reward functions, depending on the training game in question (see Table 5): these reward functions were intended to reflect the quality score computation for each game in the Clembench benchmark. Wordle (including the withclue and withcritic variants) and Referencegame used a simple 0/1 loss function: $r_g = 1$ if the agent reached the correct answer within the turn limit (six and one turns, respectively), and $r_g = 0$ otherwise.

Game	Reward Function
Taboo	$(1/len(turns)) \cdot \mathbb{I}_{corr}$
Wordle (+withclue, +withcritic)	$\mathbb{I}_{corr}$
Referencegame	$\mathbb{I}_{corr}$
Imagegame	$F_1(G_{pred}, G_{target})$
Privateshared	$Acc(S_{pred}, S_{target})$

Table 5: GRPO reward functions by training game. $\mathbb{I}_{corr} = 1$ if the agent reached the correct answer within the maximum number of turns, and $\mathbb{I}_{corr} = 0$ otherwise.

For Imagegame, r_g is given by the F_1 score between the agent's predicted grid and the target grid, and r_g is given by the agent's slot-filling accuracy for privateshared. The reward function for Taboo incorporates a length penalty: $r_g = 1/n$ if the agent guesses the correct word within $n \leq 3$ turns, and $r_g = 0$ otherwise.

For all seven games, the $r_g = 0$ if the game was aborted due to agent error, such as incorrect formatting or violation of the game's rules (e.g. predicting a non-five-letter word in Wordle).

Teacher Model. In the offline learning experiments conducted in this paper (Section 4.3.1 and Section 4.3.2), the models were trained to play both roles in all two-player games: this was not practical for the online RL experiment. If the model is trained in both roles, approximation of the optimal policy is intractable: updates to the current policy are dependent on the reward, which is dependent on the environment, which in turn is dependent on the current policy (via the generations of the current policy playing in the other role).

On the other hand, if we employ a *different*, frozen model as the second player (the *teacher* model), approximation of the optimal policy becomes tractable. However, evaluating the agent model against itself (i.e. playing both roles) introduces a mismatch between the train and test splits: the agent will have approximated the optimal policy for the environment in which the teacher is the second player.

Despite this train-test mismatch, we trained the GRPO agent with GPT-4o-mini[11] as the teacher model: in line with our goal of *learning* from interaction, we employed a more advanced model than our agent (Llama-3.1-8B), to enable the agent to learn from its teacher/caregiver. A list of the two-player games—and the roles played by the teacher and agent in each—is given in Table 9.

Challenges and Adaptations of the Playpen Environment. We adapted the Playpen environment to online RL applications by re-configuring Playpen to allow individual game instances to be played separately: this allows for the tuning of batch size as a hyperparameter, and the random permutation of game instances across batches. We additionally implemented non-agent token masking, so that the agent's loss is only computed with respect to its own generated tokens.

Teacher-Aborted Episodes In the Playpen environment, an episode can be aborted if there is a rule violation from either the agent or teacher model: for example, if the teacher model includes

[11] https://platform.openai.com/docs/models/gpt-4o-mini

the target word in its clue during a Taboo game instance.

In the case of teacher error, the agent model should not be negatively rewarded due to the aborted episode. To account for this, we set a *retry limit* ρ, such that a teacher-aborted episode will be replayed up to ρ times in the case of teacher error[12].

If a single instance $g \in G_i$ has been aborted ρ times due to teacher error, we replace g with another randomly-selected $g' \neq g \in G_i$ from the same group for loss computation and backpropagation. If *every* instance in the group G_i is aborted ρ times due to teacher error, we replace G_i with another group $G_{k \neq i}$ in the same batch.

Privateshared The privateshared game was particularly problematic for online RL, as the quality score for this game is primarily computed from probes that are conducted adjacent to the actual game, and the transcripts from these probes are removed from the agent's observations after they are completed.

Including the probes in the instance trajectory during training results in a mismatch between the train and test splits, as the agent only sees the *current* probe at test time. Conversely, removing the probes from the trajectory leads to unpredictable rewards from the environment: if the agent's reward is negatively affected by its performance in a probe, the reason for the negative reward will not be reflected in the trajectory.

For these reasons, we did not consider the agent's probing-task performance in the computation of the privateshared reward function. This has a severe negative effect on test-set performance for this game: online RL substantially degrades the model's quality score for Privateshared (see Appendix F.3), even when beginning online RL from the best SFT Llama model.

E Training details

E.1 Supervised Fine-tuning

The SFT models are fine-tuned using QLoRA (Dettmers et al., 2023) adapters ($r = 64$, $\alpha = 32$, $dropout = 0.05$) on all linear layers. The models were trained with the following arguments ($optim = adamw\_8bit$, $lr = 2e-4$, $lr\,scheduler = linear$, $decay = 0.01$, $batch\,size = 4$, $steps = 600 - 700$ and fixed $seed = 7331$). The models were quantized in 4-bit using the *unsloth* (Daniel Han and team, 2023) library and the following bits-and-bytes configuration ($use\_4bit = True$, $bnb\_4bit\_compute\_dtype = float16$, $bnb\_4bit\_quant\_type = nf4$, $use\_nested\_quant = False$). As a stopping criterion, the first checkpoint before the minimal evaluation loss that has a distance of less than or equal to 0.015 from the minimal evaluation loss was chosen. Hence, a full epoch must be trained to determine the optimal checkpoint. The most relevant libraries and their versions are ($torch = 2.4.0$, $unsloth = 2024.8$, $transformers = 4.47.1$, $bitsandbytes = 0.43.3$, $trl = 0.9.6$, $accelerate = 0.34.2$).

Training Setup All previously described experiments were conducted on a quantized version of Llama-3.1-8B (Instruct version). All models were fine-tuned using Unsloth (Daniel Han and team, 2023) with 4-bit quantization and QLoRA (Dettmers et al., 2023) for a more efficient and resource-optimized fine-tuning.

Hardware The training was conducted on one NVIDIA A100 GPU with 80 GB of VRAM and one NVIDIA H100 GPU with 95 GB of VRAM. It should be noted that technically a multi-GPU setup was possible, but every experiment was only conducted on a single-GPU.

Training Procedure In the first step, the models were trained on all available training data. Based on the training statistics (train and evaluation loss), a second model was trained using the number of steps with the lowest evaluation loss.

To address the issue of over-fitting, a third model was trained using significantly fewer steps. The number of steps was chosen based on the evaluation loss, with a threshold set to 0.015. The third model was trained until the evaluation loss reached a value within or equal to this threshold relative to the best evaluation loss.

As an example: The first model is trained for 1700 steps (all available data), but the minimal evaluation loss is reached at around 1100 steps. With a minimal eval-loss of 0.2315, the second model is trained for 1100 steps while the third model is trained until the eval-loss reaches the threshold of 0.2315 + 0.015 = 0.2430. The final model required 700 training steps. This approach helps to

[12] In practice, we set $\rho = 1$ for all experiments due to computational resource limitations.

prevent over-fitting, as continuing training beyond the threshold (where evaluation loss increases by 0.015) provides diminishing returns while potentially reducing generalization capabilities. This is a strategy we call `SFT (Cold Start, CS)` in the main paper.

We also experimented with other variants that we report below:

- `SFT (Warm Start, WS)`: Before training on the interaction data, the model was trained on 100 steps (400 samples) of synthetic instruction following tasks, using the findings from (Gupta et al., 2023) and focusing on instruction following abilities.
- `SFT (Rehearsal, R)`: during training, we interleave the gameplay training dataset with basic instruction following data following a similar approach to Lambert et al. (2025).

E.2 DPO

For both DPO Dialogue and DPO Turn, an SFT QLoRA adapter has been mounted on top of the base model Llama3.1-8B. To merge the base model and the SFT adapter, three merging strategies have been tested before DPO: merging the full-precision Llama3.1-8B model with the adapter, merging the unsloth 4-bit quantized Llama3.1-8B version with the adapter in 16bit, and merging the unsloth 4-bit quantized Llama3.1-8B with the adapter in 4bit. As reported in Tab. 6, the first strategy outperforms the others, showing comparable results to the unmerged adapter.

Model	ClemScore	pp	qs
unmerged	46.82	75.24	62.23
full-precision	47.79	74.88	63.82
16bit	33.52	70.19	47.76
4bit	30.14	60.00	50.23

Table 6: Comparison of merging strategies in terms of Clemscore, average % played (pp) and quality score (qs).

DPO training is performed on top of the 4-bit quantized SFT model, with the same bits-and-bytes configuration as the SFT models ($use\_4bit = True$, $bnb\_4bit\_compute\_dtype = float16$, $bnb\_4bit\_quant\_type = nf4$, $use\_nested\_quant = False$). QLoRA adapters are employed on the same modules as for SFT (with $r = 64$, $\alpha = 64$, and *dropout* $= 0$). The models have been trained with the $adamw\_8bit$ optimizer, a learning rate of $5e-6$, with $linear$ lr scheduler and the $beta = 0.1$ ($decay = 0$, $batch\,size = 2$, $gradient\,accumulation\,steps = 3$ and fixed $seed = 42$). During training, we evaluate the model every 20% on held-out training samples. At the end of training, only the best-performing checkpoints on the dev sets were saved. The relevant libraries' versions are: $torch = 2.5.1$, $unsloth = 2024.12.4$, $transformers = 4.46.3$, $bitsandbytes = 0.45.0$, $trl = 0.12.2$, $accelerate = 1.2.0$.

In terms of hardware, DPO development has been performed on 2xA5000s. Large differences have been observed when comparing results obtained on the A5000 and A100. The final training for clembench *v2.0* has been performed on an RTX3090 with 24GB RAM.

E.3 GRPO

We conducted two online RL experiments: one pure RL experiment, in which we initialized the agent from the baseline Llama-3.1-8B model (Instruct version) (GRPO); and a second experiment in which the RL agent was initialized from the best-performing SFT model (SFT(CS)+GRPO).

The training set for both experiments consisted of game instances from Clembench V0.9 and V1.0 for Taboo (90 instances), Wordle (60), Wordle-withclue (60), Wordle-withcritic (60), Referencegame (256), Imagegame (80), and Privateshared (80), for a total of 686 instances. The validation split consisted of 420 Clembench V1.6 game instances (total) for the training games.

Both GRPO models were trained on four NVIDIA H100 GPUs with 80 GB of VRAM (each): for speedup, trajectory generation was parallelized across the four GPUs.

Both GRPO models were tuned using LoRA (Hu et al., 2022) adapters ($r = 64$, $\alpha = 128$, *dropout* $= 0$) on their Q, K, V, and O attention projection matrices. We trained the models for five epochs on 686 game instances with a temperature of 0.75, a batch size of 16, a group size of 8, KL regularization $\beta = 0.04$, and a learn rate of 10^{-6} using the Adam optimizer (for GRPO $seed = 250329152534053703$, for SFT(CS)+GRPO $seed = 250327114458100881$).

E.4 Evaluation Details

The code for performing the experiments on the non-interactive benchmarks is available here:

https://github.com/momentino/playpen_eval/tree/playpen.

The evaluation of the models on non-interactive datasets was conducted on Ampere-architecture GPUs (A100, A40). The experiments have been conducted by extending the *lm-eval* framework with the tasks which were not present in its original version (i.e. CLadder, LM-Pragmatics, NATURAL PLAN, GLUE Diagnostics, SimpleToM). Out of these, CLadder, NATURAL PLAN and SimpleToM have been taken without any modification from those implemented by (Momentè et al., 2025). NATURAL PLAN has also been taken from there, but we removed the upper and lower bounds on the number of tokens that the model is allowed to generate. GLUE Diagnostics was implemented from scratch. For ensuring a more efficient evaluation, we relied on the vLLM library (Kwon et al., 2023) (version 0.8.3).

To ensure comparability of the results, all evaluations on clembench v2.0 were carried out exclusively on an H100 GPU. It appears that when using different GPUs, the results can differ by up to 5 percentage points in some models. The H100 was chosen due to its higher inference speeds to save time on evaluation.

F Results

F.1 Supervised Fine-tuning

Table 7 depicts results achieved by the Llama-3.1-8B-Instruct baseline and the three variants Cold Start (CS), Warm Start (WS), and Rehearsal (R) on games from `clembench` 2.0. The upper half of the table shows results for in-domain games while the lower half those for out-of-domain games.

It becomes visible that in-domain the three fine-tuned models appear to have quite substantial performance gains compared to the baseline, while for out-of-domain games the opposite is true.

Comparing the three fine-tuning versions it appears that, overall, the cold start one outperforms the other two. While WS and R show slight decreases in PrivateShared and ReferenceGame (in-domain games), the CS version displays improvements for all in-domain games.

Regarding out-of-domain games, the performance in- and decreases shifted between models but some patterns (e.g. Codenames and Map Navigation improvements) still remain. A larger discrepancy can be seen for Map Navigation (Graph). Here, WS and R seem to negatively impact the model performance. Map Navigation (Graph) is the only game that requires the model to produce a valid JSON-object. In this game in particular it is crucial to follow a strict output format since a malformed JSON leads to an aborted game.

Interestingly, Llama-3.1-8B (CS) shows out-of-domain a more substantial decrease in the % played score compared to quality score. For the other two models it appears that the performance loss is more balanced across the two scores. For Llama-3.1-8B (CS) this indicates that the fine-tuning negatively impacts the models ability to properly play the game. The % played is an indicator of what % of episodes were actually played and how many were aborted. This is tightly bound to game-specific output formats especially for the in-domain games. As for the out-of-domain games, the played score sometimes will be also negatively impacted if the model reaches a turn-limit. Even though the model knows how to structure the output, the episode will be counted as aborted. This makes it difficult to pinpoint the exact reason for the decrease in % played to one particular cause. The reason may be caused by an over-fitting to the prompt structure of the in-domain games-specific instructions. Alternatively, it could also be that other abilities such as the contextual awareness of the model are worsened by the fine-tuning process, and this may lead to reach the turn limit more easily.

F.2 DPO

Detailed results are provided in Table 8. Compared to the base L3-8B SFT(CS), both DPO Dialogue and DPO turn appear to result in a degradation of performance on in-domain games, with the most pronounced declines observed for Wordle. On the other hand, in out-of-domain games, improvements in many games are observed for DPO Dialogue, with peaks in Map Navigation (Graph) and Map Navigation (Room). DPO Turn, instead, seems to perform worse than the baseline model for most of the games.

F.3 GRPO

The performance increases and decreases for the GRPO models relative to their respective baselines are given in Table 10. Pure reinforcement learning leads to near-across-the-board improvements over the baseline Llama 3.1 8B model on all in- and out-of-domain games, although we observe slight decreases in percentage played on Wordle, Wordle With Critic, Map Navigation, Map Navi-

Game/Model	L3-8B (Baseline)	PP L3-8B CS	PP L3-8B WS	PP L3-8B R
In Domain	pp/qs	pp/qs	pp/qs	pp/qs
ImageGame	67.8/54.62	32.20/39.87	32.20/37.65	32.20/39.19
PrivateShared	100/23.48	0.00/73.65	0.00/73.05	-4.00/69.16
ReferenceGame	100/38.89	0.00/7.78	0.00/-3.33	0.00/-4.45
Taboo	98.33/31.92	1.67/5.58	1.67/3.91	1.67/9.19
Wordle	36.67/0	20.00/1.18	30.00/5.00	16.66/8.12
Wordle With Clue	0/-	23.33/71.43	10.00/16.67	6.67/0.00
Wordle With Critic	6.67/50	6.66/0.00	-3.34/50.00	-3.34/-16.67
Out-of-Domain	pp/qs	pp/qs	pp/qs	pp/qs
Adventure Game	35.94/33.85	-17.97/-18.23	-15.63/-26.93	-15.36/-17.70
Codenames	43.08/16.07	-17.70/5.14	-26.93/17.26	-17.7/5.14
Map Navigation	36/55.46	32.00/-8.75	24.00/-2.31	32.00/-0.11
Map Navigation (Graph)	20/44.33	-3.33/-7.54	-16.67/-15.76	-13.33/-13.38
Map Navigation (Room)	56.67/94.12	-16.67/-2.45	-6.67/-7.45	16.66/-16.85
MatchIt (ASCII)	100/60	-2.50/9.23	-10.00/-26.67	0.00/7.50
GuessWhat	90/11.73	-1.67/19.09	-18.33/20.05	-13.33/-3.67

Table 7: Gains and losses w.r.t baselines of average % played and quality score of individual games; L3: Llama-3.1-8B-Instruct, PP: Playpen, CS: Cold Start, WS: Warm Start, R: Rehearsal.

Game/Model	L3-8B SFT(CS)	SFT(CS)+DPO Dialogue	SFT(CS)+DPO Turn
In Domain	pp/qs	pp/qs	pp/qs
ImageGame	100/94.49	-15.25/-15.07	0.0/-1.93
PrivateShared	100/97.13	0.0/-4.40	0.0/-1.28
ReferenceGame	100/46.67	0.0/-4.45	0.0/-5.56
Taboo	100/37.5	0.0/6.94	0.0/7.22
Wordle	56.67/1.18	-30.0/2.57	-13.34/1.13
Wordle With Clue	23.33/71.43	0.0/-21.43	10.0/-48.93
Wordle With Critic	26.66/27.78	-3.33/2.78	13.34/-25.0
Out-of-Domain	pp/qs	pp/qs	pp/qs
Adventure Game	17.97/15.62	2.34/-3.9	-3.91/0.0
Codenames	25.38/21.21	-15.38/-5.83	-8.46/-7.57
Map Navigation	68/46.71	12.0/9.0	-8.0/5.28
Map Navigation (Graph)	16.67/36.79	43.33/9.18	-3.34/-0.91
Map Navigation (Room)	40/91.67	23.33/-7.46	8.33/-13.34
MatchIt (ASCII)	97.5/69.23	-70.0/-51.05	0.0/-10.26
GuessWhat	88.33/30.82	6.67/-3.92	10.0/-4.27

Table 8: Comparison of % played (pp) and quality score (qs) on individual games for the L3-8B SFT(CS) and the further trained DPO Dialogue and DPO Turn.

Game	Agent Role	Teacher Role
ImageGame	Instruction Follower	Instruction Giver
ReferenceGame	Instruction Follower	Instruction Giver
Taboo	Guesser	Describer
Wordle With Critic	Guesser	Critic

Table 9: Two-player games from the train split, and the roles played by the agent and teacher models in each for the online RL experiment.

gation (Room), and GuessWhat, along with slight decreases in quality score for the latter three out-of-domain games.

On the other hand, GRPO struggles to improve the SFT Llama model (SFT(CS)+GRPO), and only results in slight increases in quality score for Taboo, Wordle, and Wordle With Critic. We also observe substantial decreases in in-domain performance, in particular on ImageGame and PrivateShared: the decrease in PrivateShared is to be expected, as the reward function for this game is only loosely connected to the `clemscore` (as discussed in Appendix D.4). However, GRPO greatly improves the out-of-domain clemscores of the SFT model—with the notable exceptions of MatchIt (ASCII) and GuessWhat.

F.4 Evaluation on General Instruction Following Benchmarks

We report in Table 11 and Table 12 a detailed breakdown of the results obtained in the evaluation on

Game/Model	L3-8B	GRPO	L3-8B SFT(CS)	SFT(CS)+GRPO
In Domain	pp/qs	pp/qs	pp/qs	pp/qs
ImageGame	67.8/54.62	1.69/2.65	100/94.49	-24.14/-43.38
PrivateShared	100/23.48	0.0/0.69	100/97.13	0.0/-76.17
ReferenceGame	100/38.89	0.0/4.44	100/46.67	0.0/-10.0
Taboo	98.33/31.92	1.67/2.8	100/37.5	0.0/4.39
Wordle	36.67/0	-6.37/0.0	56.67/1.18	-3.34/1.94
Wordle With Clue	0/-	0.0/-	23.33/71.43	-20.0/-21.43
Wordle With Critic	6.67/50	-3.34/50.0	26.66/27.78	-16.66/50.0
Out-of-Domain	pp/qs	pp/qs	pp/qs	pp/qs
Adventure Game	35.94/33.85	19.62/11.28	17.97/15.62	28.7/20.82
Codenames	43.08/16.07	3.84/15.08	25.38/21.21	13.85/-3.56
Map Navigation	36/55.46	-6.0/-1.35	68/46.71	6.0/13.37
Map Navigation (Graph)	20/44.33	57.78/1.36	16.67/36.79	64.28/13.83
Map Navigation (Room)	56.67/94.12	-3.34/-0.37	40/91.67	10.0/1.66
MatchIt (ASCII)	100/60	0.0/2.5	97.5/69.23	2.5/-6.73
GuessWhat	90/11.73	-10.0/-1.31	88.33/30.82	-8.33/-20.4

Table 10: Gains and losses w.r.t baseline of average % played (pp) and quality score (qs) of individual games.

general instruction following tasks considered in this study.

G Qualitative Discussion

While we know that the absolute number of aborted episodes goes down from the baseline to SFT to GRPO, the distribution of reasons for those aborted episodes might change. For out-of-domain games, we investigated these reasons. Five overarching error categories were manually grouped together; the relative distributions of errors between the different models are depicted in Figure 5. The main problem for the base Llama-3.1-8B is that of exceeding the turn limit (e.g. for the Map Navigation game, continuing to loop between already visited rooms), whereas for Llama-3.1-70B problems are mostly with the answer format. This verbose behaviour is reduced by all the types of training performed. The best performing Llama-3.1-8B version is the GRPO one, achieving the lowest absolute number in terms of aborted episodes, with fewer turn limit errors and more game mechanic understanding ones in proportion. Both for Llama-3.1-8B and Llama-3.1-70B, DPO shows the highest proportion of hallucination and context-related errors, a possible signal of overfitting to the training data.

In Section 5, we show how Llama-3.1-8B trained with GRPO data is able to generalise to out-of-domain games. One of the main reasons for this is the reduced number of aborted games due to exceeding the game's turn limit (Fig. 5). Fig. 6 reports the absolute number of aborted episodes per possible aborted reasons in the Adventure Game for Llama-3.1-8B. The GRPO trained version drastically reduces the number of overall errors in the game, with around 1/4 of the original aborted episodes due to reaching the turn limit, and not reproducing the rambling errors of the SFT version ("next_action_missing"). Adventure Game, where the GRPO's higher percentage of played games (+19.62) is coupled with a higher quality score (+11.28), is a good example of the stability of these out-of-domain gains by GRPO. For Codenames, Fig. 7 reports the absolute number of aborted episodes per possible aborted reasons for Llama-3.1-8B. We observe for GRPO a reduced number of hallucinations ("Target is hallucination" and "Guess word is hallucination") compared to the base and SFT, while not decreasing in most cases compared to the baseline. A notable exception is the "Wrong number of guesses", where the GRPO model seems not to respect the number of guesses per turn required by the game. Finally, we report a Codenames episode played by all the base Llama-3.1-8B, SFT, and GRPO. As shown in Fig. 13, the base model fails due to the common error of guessing the clue word; the SFT does not encounter errors but reveals the killer word, losing the game. The GRPO, instead, is able to play the game successfully.

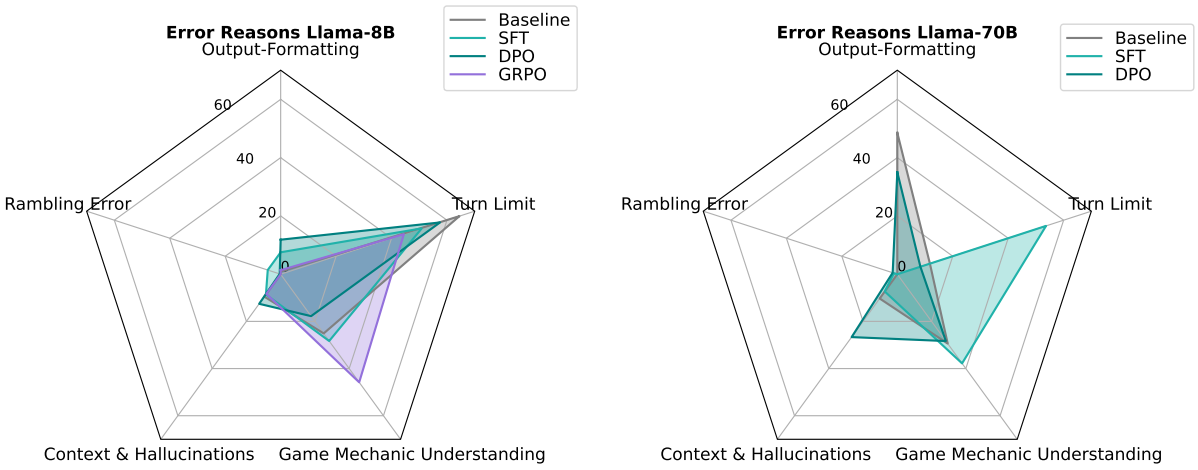

Figure 5: Relative distribution of error categories.

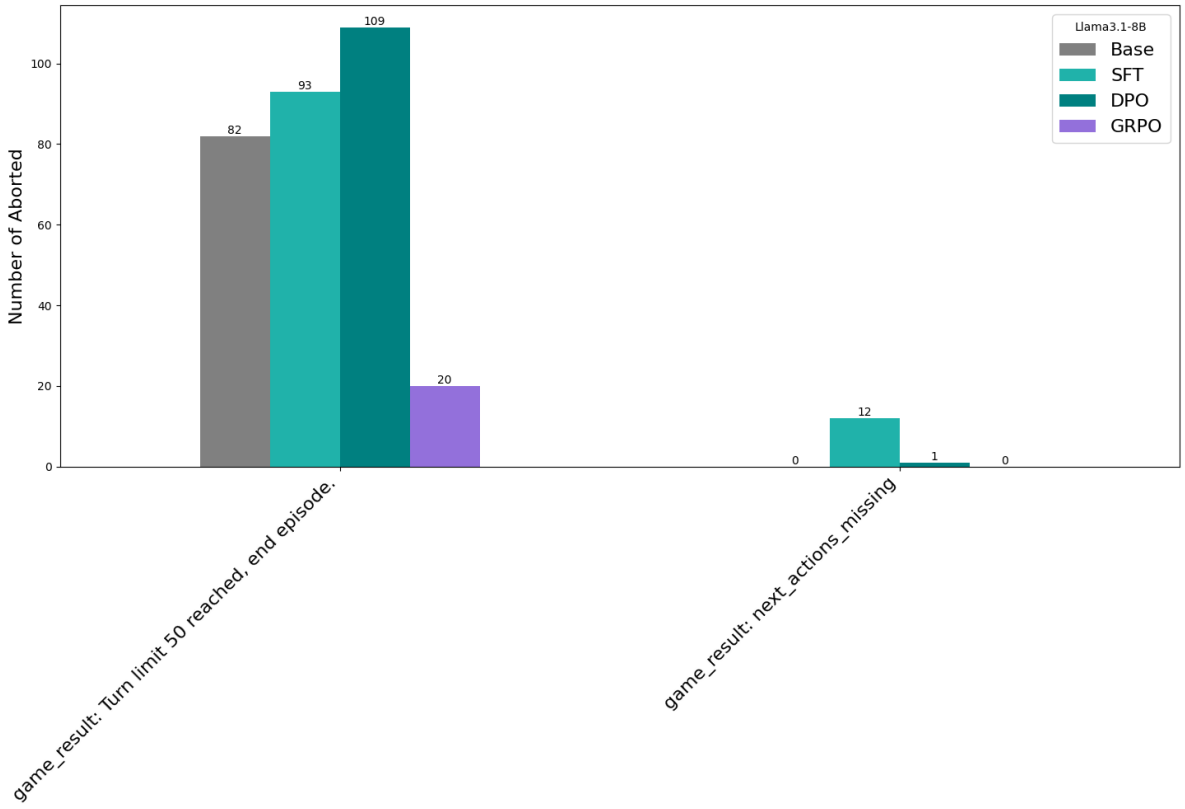

Figure 6: Number of aborted interactions per abortion reason in **Adventuregame.**

	Executive				Socio-Emotional				
	NATURAL PLAN	*LogiQA2*	*CLadder*	*WinoGrande*	*EQBench*	*LM-Pragmatics*	*SocialIQA*	*SimpleToM (AJ)*	*SimpleToM (ToM)*
Llama-3.1-8B									
Base	06.40	32.31	50.57	67.71	67.79	65.12	48.36	46.68	57.71
SFT (CS)	09.83	31.11	54.13	64.64	61.45	62.08	47.24	38.45	88.31
SFT (WS)	06.17	32.63	51.95	69.69	51.72	45.73	47.85	34.26	71.83
SFT (R)	12.80	32.18	53.37	67.88	49.84	55.97	49.84	24.80	82.39
SFT (CS) + DPO (Dial.)	12.17	27.16	53.36	61.09	61.48	50.85	44.88	38.19	80.12
SFT (CS) + DPO (Turn)	13.33	32.32	51.97	61.96	60.22	61.60	47.59	41.37	87.97
GRPO	07.31	32.12	50.96	67.17	67.69	65.49	48.56	46.60	59.20
SFT (CS) + GRPO	05.42	32.06	29.24	66.69	68.94	65.61	48.56	50.04	60.24
Llama-3.1-70B									
Base	29.03	51.52	56.34	72.77	82.03	80.97	55.02	44.33	94.5
SFT(CS)	32.03	53.24	56.95	78.45	76.40	80.61	54.96	37.84	96.43
SFT(WS)	30.75	48.72	52.39	76.60	75.50	76.58	54.86	36.01	94.59
SFT(R)	30.00	52.80	56.02	77.42	77.15	76.83	56.40	41.06	88.40
SFT(CS)+DPO(Dial.)	28.81	45.61	57.22	64.09	80.24	81.46	48.06	45.47	84.22
SFT(CS)+DPO(Turn)	28.75	48.09	56.15	67.88	81.93	83.05	52.87	48.26	84.92
Qwen-2-7B									
Base	07.40	37.21	52.99	65.51	71.42	61.22	52.25	32.17	65.82
SFT(CS)	06.94	37.66	51.64	65.43	55.61	61.10	51.64	37.10	75.33
SFT(CS)+DPO(Dial.)	6.70	35.69	51.93	62.90	57.63	64.39	49.84	32.99	79.08
SFT(CS)+DPO(Turn)	6.80	37.72	51.74	60.62	61.84	64.88	48.57	34.57	84.31

Table 11: **Performance on Executive and Socio-Emotional Tasks.** SimpleToM (AJ) and (ToM) are grouped based on the taxonomy in Momentè et al. (2025).

	Formal	**General**		**Instruction-following**
	GLUE Diagnostics	MMLU-Pro	BBH	IFEval
Llama-3.1-8B				
Base	38.06	43.35	40.37	76.88
SFT (CS)	40.23	13.16	46.75	67.25
SFT (WS)	30.74	01.70	45.52	61.40
SFT (R)	38.62	31.99	45.66	68.76
SFT (CS) + DPO (Dial.)	36.20	09.28	43.86	68.39
SFT (CS) + DPO (Turn.)	36.07	11.13	46.80	70.76
GRPO	38.68	43.73	39.31	76.97
SFT (CS) + GRPO	37.31	41.55	44.09	75.77
Llama-3.1-70B				
Base	46.16	60.37	60.74	85.16
SFT(CS)	47.72	25.90	63.91	79.38
SFT (WS)	45.86	25.03	63.58	75.10
SFT (R)	46.51	18.30	65.63	79.68
SFT(CS)+DPO(Dial.)	37.73	38.34	39.21	82.26
SFT(CS)+DPO(Turn)	39.23	36.02	53.69	85.68
Qwen-2-7B				
Base	59.68	00.42	33.24	59.16
SFT(CS)	52.98	00.18	42.53	53.87
SFT(CS)+DPO(Dial.)	53.74	00.71	29.99	54.20
SFT(CS)+DPO(Turn)	48.34	00.07	27.34	54.99

Table 12: **Model performance on formal, general and instruction-following capabilities**, as measured by GLUE Diagnostics (Wang et al., 2018), MMLU-Pro and BBH (Wang et al., 2024b; Suzgun et al., 2023), IFEval (Zhou et al., 2023b).

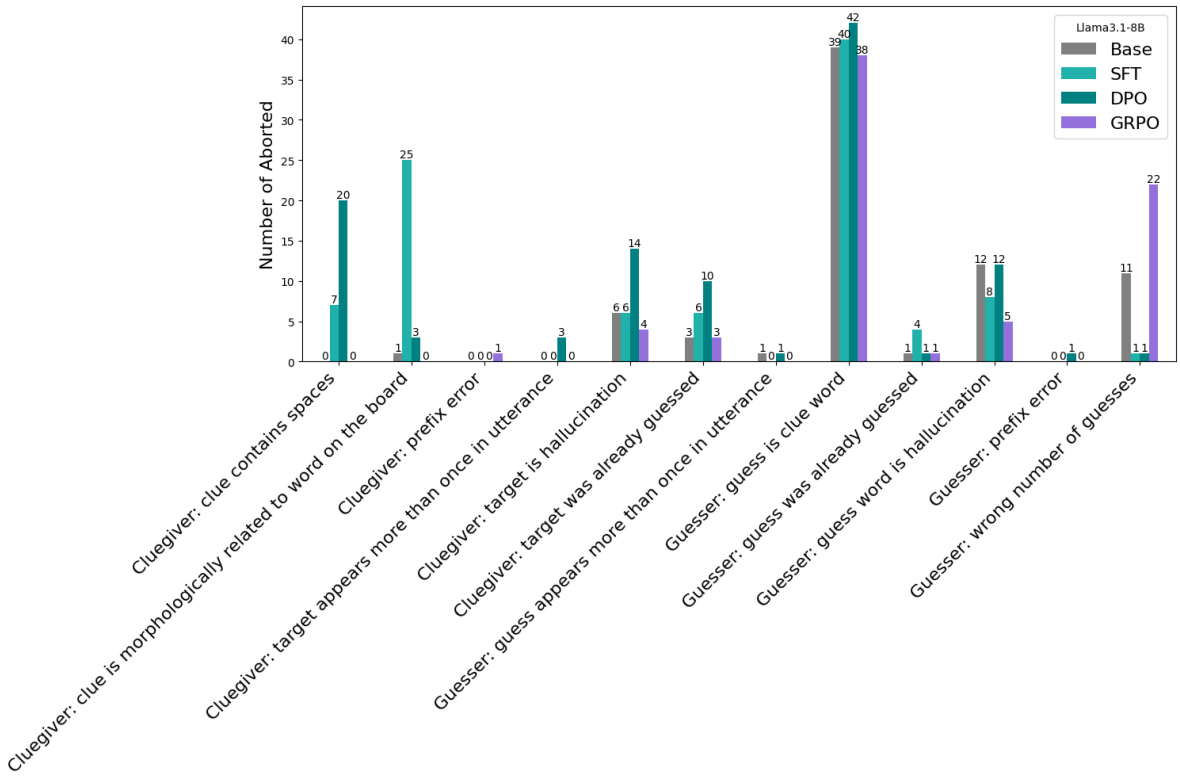

Figure 7: Number of aborted interactions per abortion reason in **Codenames**.

Table 13: Codenames sample by the Base model, the SFT and GRPO trained ones.

1	[A⟨GM] Provide one single-word clue that relates to some of your team's words from the list below. You can choose to how many and to which words your clue relates to. The goal is to have your partner guess all of your team's words based on the clues you give them, before the other team has uncovered all of theirs. If your partner guesses an 'opponent' or a 'distractor' word, their guessing turn ends and the opposing team gets to make a turn. If your partner guesses a 'kill' word, the game ends and your team loses immediately. The more words your clue relates to, the quicker you can win the game, but the harder the associations may be to guess for your partner, so choose your clue word wisely. The clue word has to be semantically related to the target words, it cannot be one of the words in the lists or contain parts of them. Always give your single-word clue and your comma-separated list of related target words in the following format and make your answers as short as possible, never include any other text than is required in this form: CLUE: <WORD> TARGETS: <TARGETS> Your team words are: gospel, shears, thunder, fog, bluegrass, attic, entrance, basement, hose. Your opponent's team words are: cotton, head, buttocks, running, back, wool, swimming, neck. Distractor words are: seasoning, brazil, oregano, china, portugal, truck, atv. The kill words are: silver.

	Base	SFT	GRPO
2	[A⟩GM] CLUE: farm TARGETS: bluegrass, shears, gospel	[A⟩GM] CLUE: roof TARGETS: attic, entrance, basement	[A⟩GM] CLUE: farm TARGETS: bluegrass, shears, gospel

Continued on next page

Table 13

#	Base	SFT	GRPO
3	[GM⟩B] Provide a comma-separated list of up to 2 words from the following list that best relate or are most closely associated with the word '**farm**'. Always start your list of guess(es) with 'GUESS: ' and do not include any other text in your answer. attic, running, neck, bluegrass, atv, brazil, portugal, seasoning, entrance, truck, china, buttocks, head, oregano, fog, swimming, wool, gospel, cotton, basement, hose, shears, back, thunder, silver	[GM⟩B] Provide a comma-separated list of up to 3 words from the following list that best relate or are most closely associated with the word 'roof'. Always start your list of guess(es) with 'GUESS: ' and do not include any other text in your answer. attic, running, neck, bluegrass, atv, brazil, portugal, seasoning, entrance, truck, china, buttocks, head, oregano, fog, swimming, wool, gospel, cotton, basement, hose, shears, back, thunder, silver	[GM⟩B] Provide a comma-separated list of up to 3 words from the following list that best relate or are most closely associated with the word 'farm'. Always start your list of guess(es) with 'GUESS: ' and do not include any other text in your answer. attic, running, neck, bluegrass, atv, brazil, portugal, seasoning, entrance, truck, china, buttocks, head, oregano, fog, swimming, wool, gospel, cotton, basement, hose, shears, back, thunder, silver
4	[GM⟨B] GUESS: wool, shears, cotton	[GM⟨B] GUESS: attic, head, back	[GM⟨B] GUESS: bluegrass, cotton, shears
5	[A⟨GM] The word wool was guessed but is an opponent word. Your teammate's turn ended there. Now provide another clue relating to some of your remaining team words and a list of the related target words. Remember to start your clue with 'CLUE: ', put a new line, and start your comma-separated list of target words with 'TARGETS: '. Notice: some words have been removed from the lists compared to previous requests. Your remaining team words are: gospel, shears, thunder, fog, bluegrass, attic, entrance, basement, hose. Remaining words for your opponent are: cotton, head, buttocks, running, back, swimming, neck. Remaining distractor words are: seasoning, brazil, oregano, china, portugal, truck, atv. Remaining kill words are: silver.	[A⟨GM] The word attic was guessed correctly. The word head was guessed but is an opponent word. Your teammate's turn ended there. Now provide another clue relating to some of your remaining team words and a list of the related target words. Remember to start your clue with 'CLUE: ', put a new line, and start your comma-separated list of target words with 'TARGETS: '. Notice: some words have been removed from the lists compared to previous requests. Your remaining team words are: gospel, shears, thunder, fog, bluegrass, entrance, basement, hose. Remaining words for your opponent are: cotton, buttocks, running, back, wool, swimming, neck. Remaining distractor words are: seasoning, brazil, oregano, china, portugal, truck, atv. Remaining kill words are: silver.	[A⟨GM] The word bluegrass was guessed correctly. The word cotton was guessed but is an opponent word. Your teammate's turn ended there. Now provide another clue relating to some of your remaining team words and a list of the related target words. Remember to start your clue with 'CLUE: ', put a new line, and start your comma-separated list of target words with 'TARGETS: '. Notice: some words have been removed from the lists compared to previous requests. Your remaining team words are: gospel, shears, thunder, fog, attic, entrance, basement, hose. Remaining words for your opponent are: head, buttocks, running, back, wool, swimming, neck. Remaining distractor words are: seasoning, brazil, oregano, china, portugal, truck, atv. Remaining kill words are: silver.
6	[A⟩GM] CLUE: storm TARGETS: thunder, fog, hose	[A⟩GM] CLUE: storm TARGETS: thunder, fog, hose	[A⟩GM] CLUE: storm TARGETS: thunder, fog

Continued on next page

Table 13

#	Base	SFT	GRPO
7	[GM⟩B] The word wool was guessed but is an opponent word. Your turn ended there. Now provide another comma-separated list of at least 1 and up to 3 words from the following list of words that best relate or are most closely associated with the word 'storm'. Remember to start your answer with 'GUESS: '. Notice: some words have been removed from the list compared to previous requests. attic, running, neck, bluegrass, atv, brazil, portugal, seasoning, entrance, truck, china, buttocks, head, oregano, fog, gospel, cotton, basement, hose, shears, back, thunder, silver	[GM⟩B] The word attic was guessed correctly. The word head was guessed but is an opponent word. Your turn ended there. Now provide another comma-separated list of at least 1 and up to 3 words from the following list of words that best relate or are most closely associated with the word 'storm'. Remember to start your answer with 'GUESS: '. Notice: some words have been removed from the list compared to previous requests. neck, bluegrass, atv, brazil, portugal, seasoning, entrance, truck, china, buttocks, oregano, fog, swimming, wool, gospel, cotton, basement, hose, shears, back, thunder, silver	[GM⟩B] The word bluegrass was guessed correctly. The word cotton was guessed but is an opponent word. Your turn ended there. Now provide another comma-separated list of at least 1 and up to 2 words from the following list of words that best relate or are most closely associated with the word 'storm'. Remember to start your answer with 'GUESS: '. Notice: some words have been removed from the list compared to previous requests. attic, running, neck, atv, brazil, portugal, seasoning, entrance, truck, china, buttocks, head, oregano, fog, swimming, wool, gospel, basement, hose, shears, thunder, silver
8	[GM⟩B] GUESS: fog, thunder, basement	[GM⟩B] GUESS: fog, thunder, back	[GM⟩B] GUESS: fog, thunder
9	[A⟨GM] The words fog, thunder, basement were guessed correctly. Your teammate's turn ended there. Now provide another clue relating to some of your remaining team words and a list of the related target words. Remember to start your clue with 'CLUE: ', put a new line, and start your comma-separated list of target words with 'TARGETS: '. Notice: some words have been removed from the lists compared to previous requests. Your remaining team words are: gospel, shears, bluegrass, attic, entrance, hose. Remaining words for your opponent are: cotton, head, buttocks, running, back, neck. Remaining distractor words are: seasoning, brazil, oregano, china, portugal, truck, atv. Remaining kill words are: silver.	[A⟨GM] The words fog, thunder were guessed correctly. The word back was guessed but is an opponent word. Your teammate's turn ended there. Now provide another clue relating to some of your remaining team words and a list of the related target words. Remember to start your clue with 'CLUE: ', put a new line, and start your comma-separated list of target words with 'TARGETS: '. Notice: some words have been removed from the lists compared to previous requests. Your remaining team words are: gospel, shears, bluegrass, entrance, basement, hose. Remaining words for your opponent are: cotton, buttocks, wool, swimming, neck. Remaining distractor words are: seasoning, brazil, oregano, china, portugal, truck, atv. Remaining kill words are: silver.	[A⟨GM] The words fog, thunder were guessed correctly. Your teammate's turn ended there. Now provide another clue relating to some of your remaining team words and a list of the related target words. Remember to start your clue with 'CLUE: ', put a new line, and start your comma-separated list of target words with 'TARGETS: '. Notice: some words have been removed from the lists compared to previous requests. Your remaining team words are: gospel, shears, attic, entrance, basement, hose. Remaining words for your opponent are: head, buttocks, running, wool, swimming, neck. Remaining distractor words are: seasoning, brazil, oregano, china, portugal, truck, atv. Remaining kill words are: silver.
10	[A⟩GM] CLUE: music TARGETS: gospel, bluegrass	[A⟩GM] CLUE: home TARGETS: entrance, basement, hose	[A⟩GM] CLUE: home TARGETS: attic, entrance, basement

Continued on next page

Table 13

#	Base	SFT	GRPO
11	[GM⟩B] The words fog, thunder, basement were guessed correctly. Your turn ended there. Now provide another comma-separated list of at least 1 and up to 2 words from the following list of words that best relate or are most closely associated with the word 'music'. Remember to start your answer with 'GUESS: '. Notice: some words have been removed from the list compared to previous requests. attic, running, neck, bluegrass, atv, brazil, portugal, seasoning, entrance, truck, china, buttocks, head, oregano, gospel, hose, shears, back, silver	[GM⟩B] The words fog, thunder were guessed correctly. The word back was guessed but is an opponent word. Your turn ended there. Now provide another comma-separated list of at least 1 and up to 3 words from the following list of words that best relate or are most closely associated with the word 'home'. Remember to start your answer with 'GUESS: '. Notice: some words have been removed from the list compared to previous requests. neck, bluegrass, atv, brazil, portugal, seasoning, entrance, truck, china, buttocks, oregano, swimming, gospel, cotton, basement, hose, shears, silver	[GM⟩B] The words fog, thunder were guessed correctly. Your turn ended there. Now provide another comma-separated list of at least 1 and up to 3 words from the following list of words that best relate or are most closely associated with the word 'home'. Remember to start your answer with 'GUESS: '. Notice: some words have been removed from the list compared to previous requests. attic, running, neck, atv, brazil, portugal, seasoning, entrance, truck, china, buttocks, oregano, swimming, wool, gospel, basement, hose, shears, silver
12	[GM⟨B] GUESS: bluegrass, gospel	[GM⟨B] GUESS: entrance, china, basement	[GM⟨B] GUESS: attic, entrance, basement
13	[A⟨GM] The words bluegrass, gospel were guessed correctly. Your teammate's turn ended there. Now provide another clue relating to some of your remaining team words and a list of the related target words. Remember to start your clue with 'CLUE: ', put a new line, and start your comma-separated list of target words with 'TARGETS: '. Notice: some words have been removed from the lists compared to previous requests. Your remaining team words are: shears, attic, entrance, hose. Remaining words for your opponent are: head, buttocks, running, back, neck. Remaining distractor words are: seasoning, brazil, oregano, china, portugal, truck, atv. Remaining kill words are: silver.	[A⟨GM] The word entrance was guessed correctly. The word china was guessed but is an innocent word. Your teammate's turn ended there. Now provide another clue relating to some of your remaining team words and a list of the related target words. Remember to start your clue with 'CLUE: ', put a new line, and start your comma-separated list of target words with 'TARGETS: '. Notice: some words have been removed from the lists compared to previous requests. Your remaining team words are: gospel, shears, bluegrass, basement, hose. Remaining words for your opponent are: cotton, buttocks, swimming, neck. Remaining distractor words are: seasoning, brazil, oregano, portugal, truck, atv. Remaining kill words are: silver.	[A⟨GM] The words attic, entrance, basement were guessed correctly. Your teammate's turn ended there. Now provide another clue relating to some of your remaining team words and a list of the related target words. Remember to start your clue with 'CLUE: ', put a new line, and start your comma-separated list of target words with 'TARGETS: '. Notice: some words have been removed from the lists compared to previous requests. Your remaining team words are: gospel, shears, hose. Remaining words for your opponent are: buttocks, running, wool, swimming, neck. Remaining distractor words are: seasoning, brazil, oregano, china, portugal, truck, atv. Remaining kill words are: silver.
14	[A⟩GM] CLUE: yard TARGETS: shears, hose	[A⟩GM] CLUE: tool TARGETS: shears, hose	[A⟩GM] CLUE: door TARGETS: entrance, attic

Continued on next page

Table 13

#	Base	SFT	GRPO
15	[GM⟩B] The words bluegrass, gospel were guessed correctly. Your turn ended there. Now provide another comma-separated list of at least 1 and up to 2 words from the following list of words that best relate or are most closely associated with the word 'door'. Remember to start your answer with 'GUESS: '. Notice: some words have been removed from the list compared to previous requests. attic, running, neck, atv, brazil, portugal, seasoning, entrance, truck, china, buttocks, oregano, hose, shears, back, silver	[GM⟩B] The word entrance was guessed correctly. The word china was guessed but is an innocent word. Your turn ended there. Now provide another comma-separated list of at least 1 and up to 2 words from the following list of words that best relate or are most closely associated with the word 'tool'. Remember to start your answer with 'GUESS: '. Notice: some words have been removed from the list compared to previous requests. neck, bluegrass, atv, brazil, portugal, seasoning, truck, oregano, swimming, gospel, cotton, basement, hose, shears, silver	[GM⟩B] The words attic, entrance, basement were guessed correctly. Your turn ended there. Now provide another comma-separated list of at least 1 and up to 2 words from the following list of words that best relate or are most closely associated with the word 'yard'. Remember to start your answer with 'GUESS: '. Notice: some words have been removed from the list compared to previous requests. running, atv, brazil, portugal, seasoning, truck, china, buttocks, oregano, swimming, wool, gospel, hose, shears, silver
16	[GM⟨B] GUESS: entrance, door	[GM⟨B] GUESS: atv, shears	[GM⟨B] GUESS: atv, hose
17	[GM|GM] Guessed word 'door' is the same word as the provided clue word, you should only select words from the provided list. [GM|GM] **Game Aborted**	[A⟨GM] The word atv was guessed but is an innocent word. Your teammate's turn ended there. Now provide another clue relating to some of your remaining team words and a list of the related target words. Remember to start your clue with 'CLUE: ', put a new line, and start your comma-separated list of target words with 'TARGETS: '. Notice: some words have been removed from the lists compared to previous requests. Your remaining team words are: gospel, shears, bluegrass, basement, hose. Remaining words for your opponent are: cotton, swimming, neck. Remaining distractor words are: seasoning, brazil, oregano, portugal, truck. Remaining kill words are: silver.	[A⟨GM] The word atv was guessed but is an innocent word. Your teammate's turn ended there. Now provide another clue relating to some of your remaining team words and a list of the related target words. Remember to start your clue with 'CLUE: ', put a new line, and start your comma-separated list of target words with 'TARGETS: '. Notice: some words have been removed from the lists compared to previous requests. Your remaining team words are: gospel, shears, hose. Remaining words for your opponent are: buttocks, running, wool, swimming. Remaining distractor words are: seasoning, brazil, oregano, china, portugal, truck. Remaining kill words are: silver.
18		[A⟩GM] CLUE: music TARGETS: gospel, bluegrass	[A⟩GM] CLUE: water TARGETS: hose

Continued on next page

Table 13

#	Base	SFT	GRPO
19		[GM⟩B] The word atv was guessed but is an innocent word. Your turn ended there. Now provide another comma-separated list of at least 1 and up to 2 words from the following list of words that best relate or are most closely associated with the word 'music'. Remember to start your answer with 'GUESS: '. Notice: some words have been removed from the list compared to previous requests. neck, bluegrass, brazil, portugal, seasoning, truck, oregano, swimming, gospel, basement, hose, shears, silver	[GM⟩B] The word atv was guessed but is an innocent word. Your turn ended there. Now provide another comma-separated list of at least 1 and up to 1 words from the following list of words that best relate or are most closely associated with the word 'water'. Remember to start your answer with 'GUESS: '. Notice: some words have been removed from the list compared to previous requests. running, brazil, portugal, seasoning, truck, china, buttocks, oregano, swimming, gospel, hose, shears, silver
20		[GM⟨B] GUESS: bluegrass, gospel	[GM⟨B] GUESS: hose
21		[A⟨GM] The words bluegrass, gospel were guessed correctly. Your teammate's turn ended there. Now provide another clue relating to some of your remaining team words and a list of the related target words. Remember to start your clue with 'CLUE: ', put a new line, and start your comma-separated list of target words with 'TARGETS: '. Notice: some words have been removed from the lists compared to previous requests. Your remaining team words are: shears, basement, hose. Remaining words for your opponent are: swimming, neck. Remaining distractor words are: seasoning, brazil, oregano, portugal, truck. Remaining kill words are: silver.	[A⟨GM] The word hose was guessed correctly. Your teammate's turn ended there. Now provide another clue relating to some of your remaining team words and a list of the related target words. Remember to start your clue with 'CLUE: ', put a new line, and start your comma-separated list of target words with 'TARGETS: '. Notice: some words have been removed from the lists compared to previous requests. Your remaining team words are: gospel, shears. Remaining words for your opponent are: buttocks, running, swimming. Remaining distractor words are: seasoning, brazil, oregano, china, portugal, truck. Remaining kill words are: silver.
22		[A⟩GM] CLUE: storage TARGETS: basement, hose	[A⟩GM] CLUE: music TARGETS: gospel

Continued on next page

Table 13

#	Base	SFT	GRPO
23		[GM⟩B] The words bluegrass, gospel were guessed correctly. Your turn ended there. Now provide another comma-separated list of at least 1 and up to 2 words from the following list of words that best relate or are most closely associated with the word 'storage'. Remember to start your answer with 'GUESS: '. Notice: some words have been removed from the list compared to previous requests. brazil, portugal, seasoning, truck, oregano, swimming, basement, hose, shears, silver	[GM⟩B] The word hose was guessed correctly. Your turn ended there. Now provide another comma-separated list of at least 1 and up to 1 words from the following list of words that best relate or are most closely associated with the word 'music'. Remember to start your answer with 'GUESS: '. Notice: some words have been removed from the list compared to previous requests. running, brazil, portugal, seasoning, truck, china, buttocks, oregano, gospel, shears, silver
24		[GM⟨B] GUESS: basement, silver	[GM⟨B] GUESS: gospel
25		[GM|GM] Team revealed. [GM|GM] **Lose**	[A⟨GM] The word gospel was guessed correctly. Your teammate's turn ended there. Now provide another clue relating to some of your remaining team words and a list of the related target words. Remember to start your clue with 'CLUE: ', put a new line, and start your comma-separated list of target words with 'TARGETS: '. Notice: some words have been removed from the lists compared to previous requests. Your remaining team words are: shears. Remaining words for your opponent are: buttocks, running. Remaining distractor words are: seasoning, brazil, oregano, china, portugal, truck. Remaining kill words are: silver.
26			[A⟩GM] CLUE: garden TARGETS: shears
27			[GM⟩B] The word gospel was guessed correctly. Your turn ended there. Now provide another comma-separated list of at least 1 and up to 1 words from the following list of words that best relate or are most closely associated with the word 'garden'. Remember to start your answer with 'GUESS: '. Notice: some words have been removed from the list compared to previous requests. brazil, portugal, seasoning, truck, china, buttocks, oregano, shears, silver

Continued on next page

Table 13

#	Base	SFT	GRPO
28			[GM⟨B] GUESS: shears
29			[GM|GM] **Game Success**