# Role-Specific Strengths and Alignment in LLM Cross-Play and Self-Play

**Kata Naszádi, Christof Monz**
Language Technology Lab, University of Amsterdam
**Correspondence:** k.naszadi@uva.nl

## Abstract

Dialogue-game benchmarks usually pair a model with itself, and the resulting self-play score is treated as a measure of that model's ability. We test whether self-play scores measure partner-independent ability by comparing them with cross-play evaluations. Across nine Clembench games, we evaluate every possible seating of three popular open-weight models. The cross-play reading exposes additional model strengths and weaknesses that are not detectable in self-play. In four of nine games, a mixed pair of two different models beats the best self-play score; in another four cases, a model does better paired with itself than paired with the strongest available partner. Looking at the transcripts, two mechanisms explain most of this: some pairs succeed because one partner writes more detailed messages that keep the interaction on track, while some self-play scores are inflated because both copies of a model make the same mistake and agree by accident.

## 1 Introduction

Dialogue games have emerged as an informative paradigm for evaluating chat-optimized language models. Rather than solving static test items, models engage in goal-directed conversations in which each generated utterance constitutes an action toward completing a shared task. Chalamalasetti et al. (2023) introduce *Clembench*, a benchmark of dialogue games spanning both asymmetric interactions, where players have distinct roles such as clue-giver and guesser, and symmetric interactions like negotiation and debate, where both players have the same role. In the standard evaluation setup, both player roles are filled by the same model. The resulting self-play score is then interpreted as a measure of that model's capability.

Recent work suggests that the self-play reading of LLM players deserves more scrutiny than it usually gets. Models recognize and favor generations resembling their own (Panickssery et al., 2024) and models increasingly make the *same* mistakes (Kim et al., 2025). The overlap in the types of mistakes grows with model capability even across model classes (Goel et al., 2025). A self-play score rewards agreement, but this agreement may stem from shared errors rather than shared competence. In language-based games, this issue is even more pronounced because language itself serves as the coordination medium. Alignment or misalignment between partners does not just occur at the level of high-level strategy, but from the way each model produces and interprets utterances.

We contribute an empirical data point to this debate: a complete cross-play and self-play evaluation of three open-weight models from different providers and architecture families (QWEN, GEMMA, MINISTRAL) across nine Clembench dialogue games. We analyze the resulting performance matrix against two prior expectations that are based on the assumption that scores reflect partner-independent competence.

**E1 (Symmetric Roles):** Self-play performance represents a baseline of capability. A more competent partner should consistently improve the joint score.

**E2 (Asymmetric Roles):** A model with a distinct strength in one role should consistently elevate its partner.

We find that E1 and E2 largely hold as baselines, and E2 reveals a split that self-play cannot expose: role-specific strengths vary across models. For example, GEMMA outperforms QWEN as the clue-giver in three separate games, while the nominally weakest model, MINISTRAL, excels as a guesser for high-frequency words because its

narrow vocabulary focus prevents it from overthinking the clues.

We also find aligned errors where successful coordination does not reflect true competence, violating E2 and creating a self-play advantage over playing with the best available partner.

## 2 Related Work

From multi-agent reinforcement learning, we know that self-play policies coordinate through private conventions that do not necessarily transfer to novel partners (Hu et al., 2020; Bard et al., 2020; Kottur et al., 2017). While such conventions are expected when policies are trained together, LLMs evaluated on dialogue games typically have no game-specific joint training. Nonetheless, modern LLMs are frequently fine-tuned on their own generated outputs (Huang et al., 2023; Wang et al., 2023). This self-training regime may implicitly result in model-specific conventions and self-preference biases, making the study of cross-play dynamics relevant to understanding how these models coordinate.

A growing number of works investigate communicative multi-agent solutions to solve complex tasks. Notably, most of these works use multiple copies of the same model to build a community of players (Wang et al., 2024; Park et al., 2023; Liang et al., 2024). An important exception is Bianchi et al. (2024), who analyzed cross-play in a negotiation setup, let different models negotiate with each other, and found that who sits across the table substantially changes the outcome, with weaker partners dragging results down.

## 3 Setup

**Models.** We use three open-weight instruction-tuned models from three providers: QWEN (Qwen3.5-27B Qwen Team, 2026), GEMMA (Gemma 4 12B-IT Gemma Team, 2026), and MINISTRAL (Ministral-3-14B-Reasoning-2512 Mistral AI, 2025). Models are invoked through Clembench's model interface with uniform greedy decoding (temperature 0.0) for every request, so any behavioral difference between models is not a sampling artifact. QWEN and MINISTRAL are reasoning-capable models. Both are run without eliciting their reasoning mode: QWEN's chat template is called with `enable_thinking=false`, MINISTRAL's reasoning channel is parsed out via a dedicated reasoning parser and is never prompted for thinking. In the tables, *qw*, *ge*, and *mi* denote Qwen3.5-27B, Gemma 4 12B-IT, and Ministral-3-14B-Reasoning-2512, respectively; the corresponding run labels are `qwen3.5-27b`, `gemma-4-12b`, and `ministral-14b`.

**Games.** We evaluate on nine Clembench dialogue games spanning two role structures. Five have *asymmetric* roles, where the two seats perform different jobs: TABOO (describer/guesser, 60 instances), REFERENCEGAME (expression-generator/answerer, 90), CODENAMES (cluegiver/guesser, 130), IMAGEGAME (instruction-leader/follower, 40), and WORDLE_WITHCRITIC (guesser/critic, 30). Four have *symmetric* roles, where both seats do the same job: CLEAN_UP (27), DOND (40), MATCHIT_ASCII (40), and HOT_AIR_BALLOON (36). In every game, P1 starts the interaction. In the symmetric games, P1 and P2 have the same role, and the labels indicate turn order. In the asymmetric games, P1 is the first role listed above—describer, expression-generator, cluegiver, instruction-leader, or guesser, respectively—and P2 is the corresponding second role. Although the roles in the symmetric games are identical, turn order can significantly affect performance (e.g., by determining which player makes the first offer in negotiation). For more detailed descriptions of the games, see Chalamalasetti et al. (2023).

**Token and context budgets.** Each game used greedy decoding and a 1024 token limit except for three games: for HOT_AIR_BALLOON, REFERENCEGAME and WORDLE_WITHCRITIC the maximum number of tokens was set to 4096. The maximum size of the context was capped at 65,536 tokens. We count any episode whose full dialogue history exceeds it as an abort with no score, on the grounds that failing to converge within the budget is itself a form of failing at the task.

**Metrics.** Each game-specific Clembench score ranges from 0 to 100. For example, WORDLE_WITHCRITIC combines success and speed, whereas REFERENCEGAME's score is its success rate (for all definitions, see Chalamalasetti et al., 2023). We distinguish between aborted episodes due to protocol violations and played episodes. We only treat a result as robust if it holds under three views: *quality*, the mean score over played episodes (blind to abort-rate differences); *clem-*

| Game | Pair | Clemscore | ΔP1 | ΔP2 |
|---|---|---|---|---|
| taboo | ge ▷ qw | 77.7 | +2.2 | +4.3 |
| codenames | ge ▷ qw | 39.2 | +11.5 | +15.4 |
| dond | qw ▷ ge | 83.3 | +36.1 | +6.9 |
| matchit_ascii | ge ▷ qw | 97.5 | +15.0 | +5.0 |

Table 1: Synergy: mixed pairs whose score beats *both* component models' own self-play score. ΔP1/ΔP2: points over Player 1's / Player 2's respective self-play.

| Game | Pair | Clemscore | ΔP1 | ΔP2 |
|---|---|---|---|---|
| dond | qw ▷ mi | 9.1 | −38.1 | −11.2 |
| matchit_ascii | qw ▷ mi | 77.5 | −15.0 | −10.0 |
| matchit_ascii | mi ▷ qw | 75.0 | −12.5 | −17.5 |
| matchit_ascii | mi ▷ ge | 70.0 | −17.5 | −12.5 |
| imagegame | qw ▷ mi | 65.1 | −31.2 | −6.8 |

Table 2: Anti-synergy: mixed pairs whose score falls *below both* component self-plays' score. dond's QWEN ▷ MINISTRAL is the matrix's strongest anomaly and reverses to 33.8 when MINISTRAL speaks first instead.

*score*, the mean score with aborted rounds entering the mean with a score of 0; and *matched instances*, quality recomputed only over the instance IDs played in *all* compared cells, which removes any selection effect from one pairing aborting on harder instances than the other. Matched-instance sample sizes range from $n$=0 to $n$=76 depending on the game and pairing. The full matrix under each of the three views is reported in Appendix A.

## 4 Results

We now present a subset of the full results, focusing on cases where pairing models from different families reveals mechanisms that self-play scores alone cannot expose. The complete $9 \times 9$ game-by-pairing matrix under all three metrics from §3 is given in Appendix A: quality (Table 4), clemscore (Table 5), and matched instances (Table 6).

First, we mention that three games in particular carry high abort rates that are worth explaining before the case studies. In CODENAMES, QWENas cluegiver drives abort rates of 43–52/130 episodes; this is because Qwen's broad clues make the receivers guess outside of the valid candidates. In HOT_AIR_BALLOON, aborts come down to: QWEN's free-form reasoning buries the required structured tags, and all 24 complexity/negotiation episodes abort in every QWEN-containing run. In IMAGEGAME GEMMAhas a high abort rate as instruction-giver because it echoes the instruction template itself rather than producing a drawing command.

In Table 1, we display cases where cross-play performance exceeds both self-play scores. This allows us to address an interesting question: do models design utterances for themselves? While there is no explicit theory-of-mind mechanism implemented in these models, we can still ask whether a model is best at resolving clues generated by its own copy. Surprisingly, the answer is no. QWEN tends to give high-level, abstract clues, which is best attested by TABOO statistics: on average, it addresses 2.4 words at once, while Gemma only targets 1.7. QWEN gets confused by the abstract clues of its own copy and often hallucinates invalid answers, yet is much more successful when paired with Gemma's more concrete strategy.

In MATCHIT_ASCII, where players have to describe ASCII-grid patterns, the same GEMMA-as-describer advantage holds over QWEN self-play. GEMMA's approach is similarly low-level as for taboo and codenames: it describes the grid cell by cell while QWEN's high-level natural language descriptions often contain inaccuracies and get misunderstood (further examples in Appendix B).

In the negotiation game "dond", QWEN typically negotiates successfully with its own copy in text, but the player making the closing final "[Proposal: ...]" botches the deal by proposing the partner's share contradicting the discussed agreement. The pairing QWEN ▷ GEMMA combines QWEN's negotiation with GEMMA's discipline of the formatted proposal generation, which results in the best performance.

In Table 2 we display cases where self-play performance exceeds both cross-play scores even though the partners exhibit stronger performance individually. These cases contradict our initial expectations (E1 and E2), under which being paired with a stronger partner should improve performance. MINISTRAL playing MATCHIT_ASCII is the strongest case for this; the self-play advantage exists purely because both copies of MINISTRAL misread the ASCII grids in the exact same way, agreeing that two grids match for entirely wrong reasons. Similar aligned misinterpretations can be found in the imagegame for MINISTRAL's self-play, but the models' unrelated errors mask this self-play advantage (see Figure 1 for an example and Appendix B for more). These are clear examples of correlated error masquerading as success-

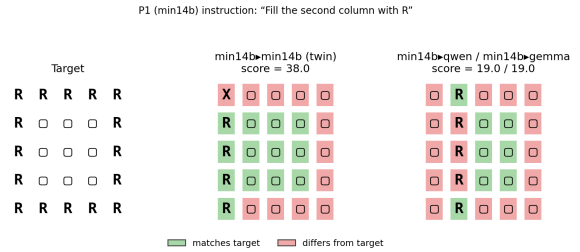


Figure 1: MINISTRAL uses "second column" to refer to the first one and its own copy decodes it the same way while GEMMA gets misled.

| | GEMMA | QWEN | MINISTRAL |
|---|---|---|---|
| high_en | 65.8 | 65.3 | **69.7** |
| medium_en | 55.6 | **57.2** | 51.4 |
| low_en | 57.2 | **66.1** | 57.2 |

Table 3: Average taboo score in the guesser role, by target-word frequency band, averaged over both partners and self-play (aborts scored 0).

ful communication.

The QWEN-MINISTRAL pairing in imagegame is another example of audience mismatch: the lower-capability MINISTRALfails with QWENas the latter gives highly compressed descriptions that only the higher-capability GEMMAcan resolve.

**Frequency effect of target words** Not every partner effect surfaces at the game level. MINISTRAL is the best model in the guessing role in taboo's high-frequency range (Table 3: 69.7 vs. 65.8 for GEMMA, 65.3 for QWEN), plausibly because it is not thrown off by assigning high probability to rarer words. The effect is most striking in its best pairing: guessing for QWEN's descriptions, MINISTRALscores 95, 15 points above Qwen's own self-play score of 80 in the same high-frequency range.

## 5 Discussion

Our evaluation set out to investigate the existence of unexpected, systemic self-play advantage in dialogue games played by LLMs. We specifically looked for dog-whistle-like patterns: a word or phrasing whose success with a given partner could not be derived from the optimal communicative strategy the game itself demands.

We find clear evidence of such a self-play advantage in only one specific condition with our weakest tested model MINISTRAL. When the game involves an external *grounding element* (such as ASCII grids), the MINISTRAL self-play exhibits consistent, correlated errors. Because both copies of MINISTRAL consistently misinterpret the grounding in the exact same way, communication succeeds. This coordination falls apart immediately in cross-play because a partner model (like GEMMA) interprets the grounded reference correctly, exposing the unconventional language use of MINISTRAL.

Most patterns reported in this paper, however, are readily interpretable within each game's dynamics and are well explained by known communicative strategies without having to turn to deeper model-specific anomalies. The best information-givers are highly specific: GEMMA's Codenames clues stay close to a small set of candidates, and its ImageGame instructions enumerate every cell rather than compressing them into a single command. The best information-receivers are literal, avoiding over-interpretation: MINISTRAL's Taboo guesses follow the common reading of a clue rather than searching for a more abstract alternative, allowing it to outperform stronger partners on high-frequency words. These differences reflect *audience design*: a speaker's optimal message depends on the addressee's ability to reconstruct what is left implicit (Naszadi et al., 2024). QWEN's REFERENCEGAME instructions are efficient, but only for a capable listener. GEMMA's longer, exhaustive descriptions are safer but less natural. Because Clembench rewards communicative accuracy more than efficiency, this cautious style is consistently favored.

The absence of stronger evidence for systemic self-play advantage does not mean such effects do not exist. Rather, the communication games considered here may not demand sufficiently sophisticated language use for detecting such conventions. In many settings, performance is still limited by more basic communicative failures. For example, in CODENAMES, hallucinated associations dominate error more than any partner-specific clue strategy would. If lower-level failures dominate the error profile, there is little room to detect more subtle "secret handshake" patterns.

## Limitations

We follow the standard *Clembench* evaluation protocol, including its recommended deterministic decoding setup (greedy decoding with temperature

0). This means that each reported score is based on a single realization of each game. Our analysis does not capture the variability that may arise under stochastic decoding. In addition, some benchmark games contain only a small number of evaluation instances, which limits the statistical power of comparisons on individual games. Finally, our analysis of language use relies on qualitative human annotation. Consequently, some factors contributing to successful or unsuccessful coordination may not be reflected in our analysis or may not align with the interpretations presented in the discussion (Section 5).

## Acknowledgments

This research was funded in part by the Netherlands Organization for Scientific Research (NWO) under project number VI.C.192.080. We also received funding from the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research with grant number 024.004.022.

## References

Nolan Bard, Jakob N Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, et al. 2020. The hanabi challenge: A new frontier for ai research. *Artificial Intelligence*, 280:103216.

Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. 2024. How well can llms negotiate? negotiationarena platform and analysis. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

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.

Gemma Team. 2026. Gemma 4 12b-it. https://huggingface.co/google/gemma-4-12B-it. Model card.

Shashwat Goel, Joschka Strüber, Ilze Amanda Auzina, Karuna K Chandra, Ponnurangam Kumaraguru, Douwe Kiela, Ameya Prabhu, Matthias Bethge, and Jonas Geiping. 2025. Great models think alike and this undermines ai oversight. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org.

Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. 2020. "other-play" for zero-shot coordination. In *International conference on machine learning*, pages 4399–4410. PMLR.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1068, Singapore. Association for Computational Linguistics.

Elliot Kim, Avi Garg, Kenny Peng, and Nikhil Garg. 2025. Correlated errors in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org.

Satwik Kottur, José Moura, Stefan Lee, and Dhruv Batra. 2017. Natural language does not emerge 'naturally' in multi-agent dialog. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2962–2967, Copenhagen, Denmark. Association for Computational Linguistics.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.

Mistral AI. 2025. Ministral-3-14b-reasoning-2512. https://huggingface.co/mistralai/Ministral-3-14B-Reasoning-2512. Model card.

Kata Naszadi, Frans A Oliehoek, and Christof Monz. 2024. Communicating with speakers and listeners of different pragmatic levels. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21777–21783, Miami, Florida, USA. Association for Computational Linguistics.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Qwen Team. 2026. Qwen3.5-27b. https://huggingface.co/Qwen/Qwen3.5-27B. Model card.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 257–279, Mexico City, Mexico. Association for Computational Linguistics.

## A Full result matrices

Tables 4–6 report the complete $9 \times 9$ game-by-pairing matrix under the three metrics from §3: Table 4 gives quality (abort-blind mean Main Score over played episodes), Table 5 gives clem-score (the same means with aborted episodes scored 0), and Table 6 gives the matched-instance view (quality recomputed only over instance IDs completed by all nine pairings). Blank cells mark pairings or games with zero completed or zero matched episodes.

## B Transcript excerpts

**MATCHIT_ASCII with MINISTRAL** Figures 2 and 3 give examples of correlated errors by MINISTRAL.

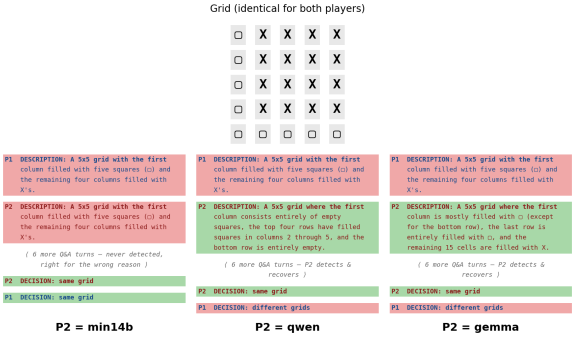


Figure 2: Two MINISTRAL instances give the same incorrect description of a grid and agree that the grids match. The error is exposed in cross-play, where the partner detects it in the dialogue.

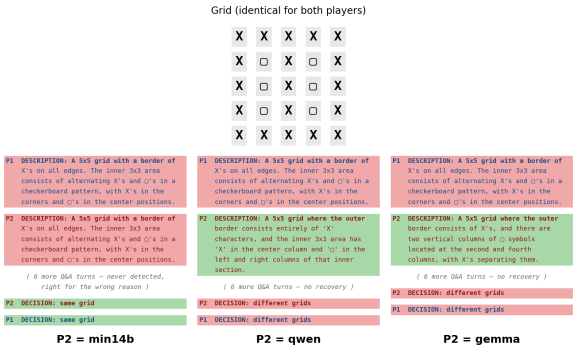


Figure 3: Two MINISTRAL instances give the same incorrect description of a grid and agree that the grids match. The other model does not detect the incorrect description, and the interaction fails for both partners.

**GEMMA advantage over QWEN as P1** Figures 4–5 give two REFERENCEGAME instances where GEMMA's detailed, enumerative expression lets MINISTRAL pick the target while QWEN's compressed expression sends MINISTRAL to a distractor. Figures 6 and 7 show two further examples contrasting MINISTRAL and GEMMA in the receiver role.

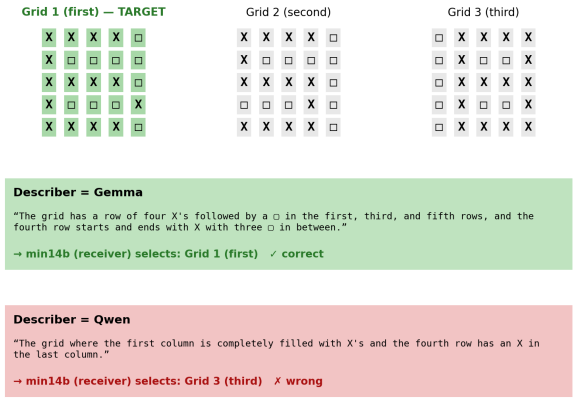


Figure 4: REFERENCEGAME instance 15 (number_grids), score 100 vs. 0, receiver fixed at MINISTRAL. Gemma describes every row. Qwen compresses to two features.

| Game | qw·qw | ge·ge | mi·mi | qw▷ge | ge▷qw | ge▷mi | qw▷mi | mi▷qw | mi▷ge |
|---|---|---|---|---|---|---|---|---|---|
| taboo | 78.6 (56) | 79.5 (57) | 43.0 (50) | 77.0 (55) | 83.3 (56) | 73.0 (58) | 75.7 (57) | 48.9 (46) | 42.4 (46) |
| referencegame | 95.6 (90) | 100.0 (87) | 62.2 (90) | 95.4 (87) | 92.2 (90) | 80.0 (90) | 68.9 (90) | 78.9 (90) | 81.6 (76) |
| codenames | 35.6 (87) | 33.6 (107) | 19.2 (78) | 34.5 (84) | 44.7 (114) | 36.1 (97) | 35.9 (78) | 26.9 (104) | 24.5 (110) |
| clean_up | 71.3 (25) | 79.4 (13) | – (0) | 75.1 (14) | 70.7 (19) | – (0) | – (0) | 59.7 (10) | 51.0 (8) |
| dond | 47.2 (40) | 76.4 (40) | 20.8 (39) | 83.3 (40) | 78.2 (40) | 35.2 (40) | 9.1 (40) | 33.8 (40) | 30.5 (39) |
| matchit_ascii | 92.5 (40) | 82.5 (40) | 87.5 (40) | 85.0 (40) | 97.5 (40) | 85.0 (40) | 77.5 (40) | 75.0 (40) | 70.0 (40) |
| imagegame | 96.3 (40) | – (0) | 73.7 (39) | 96.6 (40) | – (0) | – (0) | 66.8 (39) | 79.5 (40) | 79.8 (40) |
| hot_air_balloon | 81.1 (10) | 75.9 (29) | 59.8 (6) | 54.1 (12) | 73.8 (12) | 50.9 (21) | 64.6 (12) | 0.0 (1) | 21.0 (4) |
| wordle_withcritic | 85.4 (16) | 91.0 (13) | 75.0 (4) | 80.0 (15) | 82.8 (17) | 76.1 (15) | 75.9 (18) | 30.0 (5) | 33.3 (3) |

Table 4: Quality (mean Main Score over played episodes; played count in parentheses). This is the abort-blind view described in §3: it isolates skill on completed episodes but does not penalize a pairing for aborting more often than another. All cells 493/493 episodes finished-or-aborted; context-overflow episodes counted as aborted (103 documented stubs). qw = qwen3.5-27b, ge = gemma-4-12b, mi = ministral-14b; X▷Y: X speaks first.

| Game | qw·qw | ge·ge | mi·mi | qw▷ge | ge▷qw | ge▷mi | qw▷mi | mi▷qw | mi▷ge |
|---|---|---|---|---|---|---|---|---|---|
| taboo | 73.4 | 75.5 | 35.8 | 70.6 | 77.7 | 70.6 | 71.9 | 37.5 | 32.5 |
| referencegame | 95.6 | 96.7 | 62.2 | 92.2 | 92.2 | 80.0 | 68.9 | 78.9 | 68.9 |
| codenames | 23.8 | 27.7 | 11.5 | 22.3 | 39.2 | 26.9 | 21.5 | 21.5 | 20.7 |
| clean_up | 66.0 | 38.2 | 0.0 | 38.9 | 49.8 | 0.0 | 0.0 | 22.1 | 15.1 |
| dond | 47.2 | 76.4 | 20.3 | 83.3 | 78.2 | 35.2 | 9.1 | 33.8 | 29.7 |
| matchit_ascii | 92.5 | 82.5 | 87.5 | 85.0 | 97.5 | 85.0 | 77.5 | 75.0 | 70.0 |
| imagegame | 96.3 | 0.0 | 71.9 | 96.6 | 0.0 | 0.0 | 65.1 | 79.5 | 79.8 |
| hot_air_balloon | 22.5 | 61.1 | 10.0 | 18.0 | 24.6 | 29.7 | 21.5 | 0.0 | 2.3 |
| wordle_withcritic | 45.5 | 39.4 | 10.0 | 40.0 | 46.9 | 38.0 | 45.5 | 5.0 | 3.3 |

Table 5: Clemscore (aborted episodes scored 0). This is the abort-sensitive counterpart to Table 4: it folds a pairing's abort rate back into the mean, so the two tables read together separate genuine skill differences from differences driven purely by protocol failures. qw = qwen3.5-27b, ge = gemma-4-12b, mi = ministral-14b; X▷Y: X speaks first.

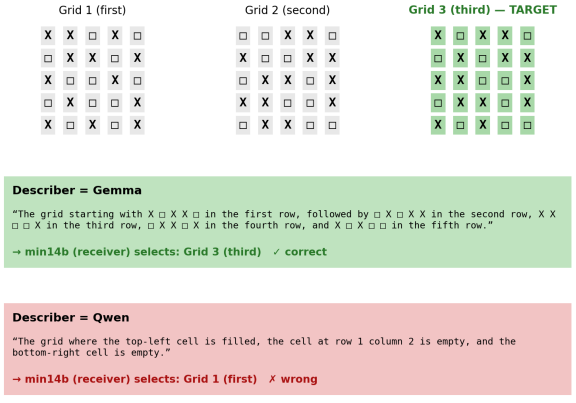


Figure 5: REFERENCEGAME instance 11 (`random_grids`), score 100 vs. 0, receiver fixed at MINISTRAL. GEMMAtranscribes the pattern row by row, giving MINISTRAL the full target; Qwen names only three distinguishing cells and MINISTRAL selects the wrong one.

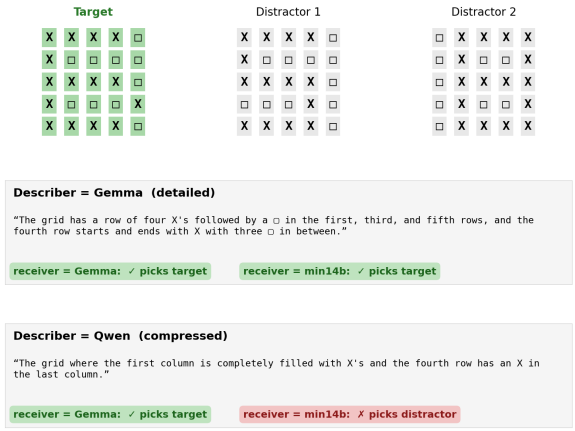


Figure 6: REFERENCEGAME instance 15 (`number_grids`). REFERENCEGAME is single-shot, so each describer's referring expression is identical across receivers. GEMMA's detailed, row-by-row expression is understood by both receivers. QWEN's two-feature expression is enough for the stronger GEMMA receiver to recover the target but not for MINISTRAL, which selects a distractor.

| Game | $n$ | qw·qw | ge·ge | mi·mi | qw▷ge | ge▷qw | ge▷mi | qw▷mi | mi▷qw | mi▷ge |
|---|---|---|---|---|---|---|---|---|---|---|
| taboo | 41 | 82.9 | 85.8 | 45.1 | 81.7 | 89.8 | 81.7 | 82.5 | 46.3 | 41.9 |
| referencegame | 76 | 96.1 | 100.0 | 65.8 | 100.0 | 93.4 | 81.6 | 72.4 | 81.6 | 81.6 |
| codenames | 16 | 50.0 | 56.2 | 18.8 | 56.2 | 56.2 | 31.2 | 56.2 | 37.5 | 31.2 |
| clean_up | 2 | 94.8 | 100.0 | – | 88.1 | 100.0 | – | – | 39.7 | 53.7 |
| dond | 38 | 49.7 | 77.8 | 20.8 | 83.0 | 79.7 | 37.0 | 9.5 | 32.4 | 31.3 |
| matchit_ascii | 40 | 92.5 | 82.5 | 87.5 | 85.0 | 97.5 | 85.0 | 77.5 | 75.0 | 70.0 |
| imagegame | 38 | 96.1 | – | 73.0 | 96.4 | – | – | 66.4 | 78.6 | 78.9 |
| hot_air_balloon | 0 | – | – | – | – | – | – | – | – | – |
| wordle_withcritic | 0 | – | – | – | – | – | – | – | – | – |

Table 6: Matched instances: mean Main Score restricted to the instance IDs played (non-aborted) in every compared pairing for a given game, removing any selection effect from one pairing aborting on harder instances than another (§3). $n$ gives the number of matched instance IDs; games where no instance is completed by all nine pairings (HOT_AIR_BALLOON, WORDLE_WITHCRITIC) are left blank.

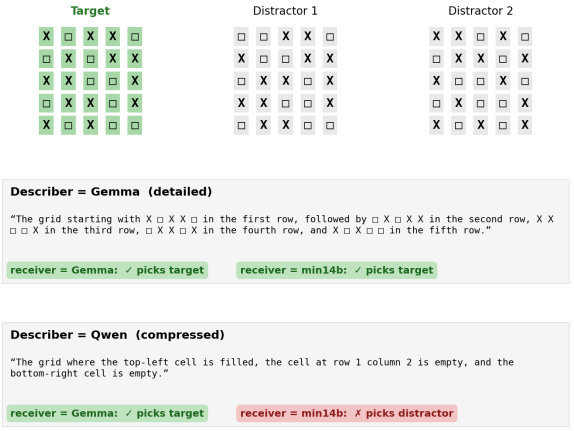


Figure 7: REFERENCEGAME instance 11 (`random_grids`), same describer × receiver design as Figure 6. On an irregular grid GEMMA transcribes the pattern row by row and both receivers succeed. Qwen names only three distinguishing cells.