MeepleLM: A Virtual Playtester Simulating Diverse Subjective Experiences

Anonymous ACL submission

Abstract

Recent advancements have expanded the role of Large Language Models in board games from playing agents to creative co-designers. However, a critical gap remains: current systems lack the capacity to offer constructive critique grounded in the emergent user experience. Bridging this gap is fundamental for harmonizing *Human-AI collaboration*, as it empowers designers to refine their creations via external perspectives while steering models away from biased or unpredictable outcomes. Automating critique for board games presents two challenges: inferring the *latent dynamics* connecting static rules to gameplay without an explicit engine, and modeling the *subjective heterogeneity* of diverse player groups. To address these, we curate a comprehensive dataset of 1,727 structurally corrected rulebooks and 150K reviews selected via rigorous quality scoring and facet-aware sampling. We augment this data with *Mechanics-Dynamics-Aesthetics (MDA)* reasoning to explicitly bridge the causal gap between written rules and player experience. We further distill distinct player personas and introduce **MeepleLM**, a specialized model that internalizes persona-specific reasoning patterns to accurately simulate the subjective feedback of diverse player archetypes. Extensive experiments demonstrate that MeepleLM significantly outperforms latest commercial models (e.g., GPT-5.1, Gemini3-Pro) in community alignment and critique quality, achieving a 70% preference rate in user studies assessing practical utility. Ultimately, MeepleLM serves as a reliable *virtual playtester* for general interactive systems, marking a pivotal step towards audience-aligned, experience-aware Human-AI collaboration.

1 Introduction

Board games have long served as a universal medium of significant cultural and economic value, captivating a vast global audience (Rodríguez, 2025). Recently, the rapid development of Large Language Models (LLMs) has introduced unprecedented possibilities to this classic domain. Specifically, board games serve as a prominent platform for evaluating diverse model facets, ranging from reasoning (Lin et al., 2025) and decision-making (Tang et al., 2025) to role-playing (Yu et al., 2025) and social simulation (Hansteen Izora and Teuscher, 2025). Beyond serving as a testbed, recent research highlights board game development as a pivotal domain for investigating *Human-AI Collaboration*, where LLMs serve as active co-designers to perform tasks such as generating mechanics (Patrick and Khan, 2025), facilitating iterative prototyping (Ma et al., 2025) and synthesizing executable engines (Hong et al., 2025; Lehrach et al., 2025).

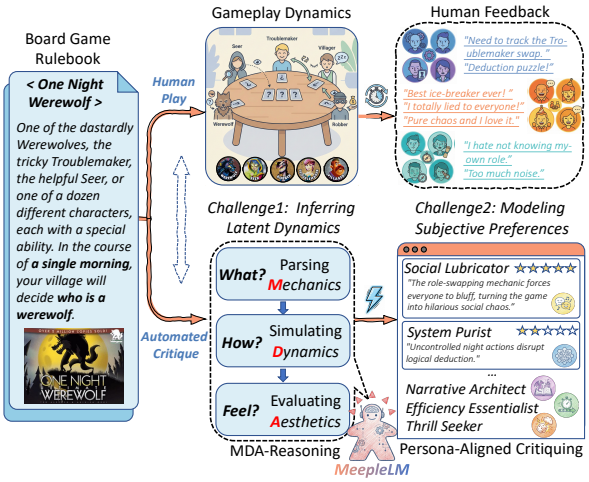


Figure 1: **Overview of MeepleLM.** Acting as a *Virtual Playtester*, the model offers a rapid, automated alternative to the resource-intensive Human Play loop. By leveraging MDA-Reasoning to infer latent dynamics from Static Rulebooks, MeepleLM generates Persona-Aligned Critiques tailored to diverse player archetypes.

However, while automated development has advanced, a critical gap remains: current systems lack the capacity to offer constructive critique grounded

in the emergent user experience. Such feedback is vital for harmonizing the roles of both human and model in the co-creation loop. On one hand, designers require external perspectives to refine their crations and better comprehend their audience (Fang et al., 2025), a process that ultimately catalyzes further creativity (Choi et al., 2025). On the other hand, for LLM-driven systems, the absence of effective user feedback can lead to biased content (Taveekitworachai et al., 2024) or unpredictable, frustrating experiences (Yong and Mitchell, 2023). Consequently, bridging this gap is fundamental to evolving LLMs into empathetic partners for *Human-AI Collaboration*. By prioritizing diverse user experiences, this approach ensures that future co-creation is driven not merely by technical validity, but by a dynamic alignment with individual needs.

To this end, we need an evaluation paradigm that can map "design intent" (or "system specifications") to "user experience." However, board game experiences are characteristically emergent and subjective: they are not static properties of rulebooks, but are jointly generated through interaction as mechanics unfold, players MDA, and emotional responses arise (Forlizzi and Battarbee, 2004). This inherent characteristic poses two core challenges for automated evaluation: (1) **Inferring Latent Dynamics from Static Rules.** While rulebooks serve as explicit "code", gameplay experience is an emergent property generated only when mechanics interact at runtime. The core challenge is to bridge the gap between written specifications and dynamic interactions. Since LLMs lack an explicit game engine, they must infer plausible execution trajectories from rules and use empirical player feedback as an external signal to recover latent causal links that connect mechanics to outcomes and reactions. (2) **Modeling Subjective Group Preferences.** Experience is not universal; the same mechanism can elicit conflicting reactions across different player demographics (e.g., high randomness may delight a Socializer but frustrate a Strategist). If critiques collapse into an average "one-size-fits-all" judgment, they become generic and less actionable for design or recommendation. The challenge, therefore, is to model this subjective heterogeneity by aligning reasoning with specific group preferences, simulating distinct personas rather than a single "standard" user.

To address these challenges, we meticulously curate a large-scale dataset of structurally corrected rulebooks from selected board games, paired with reviews filtered through rigorous scoring and quality assessment. We further augment this data by incorporating the classic game design theory of Mechanics-Dynamics-Aesthetics (MDA) (Hunicke et al., 2004) into Chain-of-Thought (CoT) reasoning, thereby making the latent execution logic explicit. To structure the inherent subjectivity of feedback, we distill distinct player personas through an expert–LLM collaborative interpretation of data-driven community clusters. Building upon this foundation, we introduce MeepleLM(Figure 1), a specialized model designed to predict gameplay experiences by simulating the perspectives of real-world players. Extensive experiments validate that MeepleLM significantly outperforms state-of-the-art baselines in capturing authentic user experiences.

Our contributions are summarized as follows:

- We present the first systematic study on the **automated evaluation of board games**. We bridge the gap between static rules and distinct player experiences by simulating the latent gameplay dynamics.
- We curate a high-quality dataset of **1,727 rulebooks** and **150K critiques**, selected via rigorous filtering and quality evaluation. We further leverage the MDA framework to synthesize explicit COT paths that recover the latent dynamics connecting rules to experiences.
- We distill five data-driven player personas and introduce **MeepleLM**. By internalizing persona-specific reasoning, our model predicts authentic gameplay experiences that reflect the diverse preferences of real-world communities.
- We conduct a systematic evaluation on a stratified set of **207 games**. Experiments across macro-level alignment, micro-level fidelity, and practical utility demonstrate that MeepleLM significantly outperforms state-of-the-art LLMs as a reliable virtual playtester.

Ultimately, by bridging static rules and dynamic experiences, MeepleLM establishes a paradigm for the **automated virtual testing of general interactive systems**, which accelerates design iteration via anticipated market feedback and facilitates personalized selection for players. This paves the way for **experience-aware Human-AI collaboration**,

where models evolve from functional tools into empathetic partners attuned to **subjective audience sensibilities**.

2 Related Work

LLM-Driven Feedback and Assistance. Recent advancements have empowered LLMs to surpass traditional metrics in evaluating open-ended text, demonstrating high alignment with human judgments (Li et al., 2024; Gao et al., 2025; Chen et al., 2023). Current systems provide constructive feedback ranging from granular writing issues (Russell et al., 2025) to comprehensive peer reviews (Benharrak et al., 2024; Rashkin et al., 2025), with established metrics for narrative consistency (Rashkin et al., 2025), structural integrity (Zheng et al., 2025), and subjective enjoyment (Yang and Jin, 2025). However, these approaches treat text as *static narratives*, failing to address the *executable logic* inherent in interactive systems like board games. Even when extended to design assistance, such as generating levels (Todd et al., 2023) or rule codes (Todd et al., 2024; Tanaka and Simo-Serra, 2024), LLMs often prioritize syntactic correctness over logical coherence. As noted in recent studies, this frequently yields functional yet meaningless mechanics, known as *introns* (Todd et al., 2024; Hu et al., 2024). While existing tools fragment into abstract brainstorming (Li et al., 2025) or rigid asset production (Lindfors, 2025) without assessing playability, our work bridges this gap by simulating *dynamic interactions* to predict the *emergent experience* directly from rules.

User Simulation and Persona Modeling. Understanding audience heterogeneity is crucial for creators, yet manual analysis of diverse feedback is cognitively demanding (Choi et al., 2023; Ma et al., 2023). The field has evolved from survey-based "imagined users" (Cooper, 1999; Salminen et al., 2018) to algorithmic clustering (McGinn and Kotamraju, 2008; Salminen et al., 2020) and, recently, LLM-based perspective simulation (Benharrak et al., 2024; Park et al., 2023). However, purely synthetic simulations often lack ecological validity, risking hallucinations or stereotyping due to foundation model biases (Cheng et al., 2023). To mitigate this, state-of-the-art systems emphasize grounding simulations in empirical data for *representativeness* (Shin et al., 2024; Choi et al., 2025). Aligning with this paradigm, we ground personas in large-scale gameplay critiques rather than conversational history. This allows our model to internalize distinct preferences, facilitating *Human-AI collaboration* through empathetic, persona-aligned feedback rather than generic judgments.

3 Data Construction

We curated a multi-layered dataset that maps objective game rulebooks to subjective player feedback across diverse personas. The overall construction pipeline is illustrated in Figure 2.

3.1 Game Selection

We curated a collection of **1,727 board games** via a stratified sampling strategy on BoardGameGeek (BGG)[1]. As detailed in Appendix A, our selection prioritizes four dimensions to ensure a comprehensive representation of the design landscape: (1) **Market Stratification:** To mitigate survivorship bias, we balanced the selection between "elite" titles (including 83 from the Top 100) and "long-tail" designs (comprising 53% of the dataset with Rank $> 1,000$), capturing the full spectrum of market reception. (2) **Cognitive Spectrum:** We covered the entire range of BGG Weight (1.0–5.0), encompassing everything from low-burden social party games to calculation-intensive strategy simulations. (3) **Temporal Span:** The dataset balances historical depth with modern relevance, containing 47% classic titles released pre-2015 alongside 35 cutting-edge designs from 2024 and beyond. (4) **Mechanical Heterogeneity:** To ensure structural diversity, the collection spans 192 unique mechanics and 81 themes, covering logic distinct from standard genres.

3.2 Rulebook Structuring

We processed the official rulebooks into a structured knowledge base via a three-step pipeline. First, we parsed raw PDFs into hierarchical Markdown using Mineru (Niu et al., 2025) to preserve layout information. Second, we prompted Qwen3-235B(Yang et al., 2025) to restructure the raw text into a standardized hierarchical format (e.g., unifying headers for *Objective*, *Components*, and *Flow*); the specific extraction prompt is detailed in Appendix B.1, and a complete structured example is provided in Appendix B.2. Finally, to ensure maximum fidelity, we employed GPT-5.1 to cross-reference and rectify the initial drafts against the source text, correcting logical inconsistencies or

[1] https://boardgamegeek.com/browse/boardgame

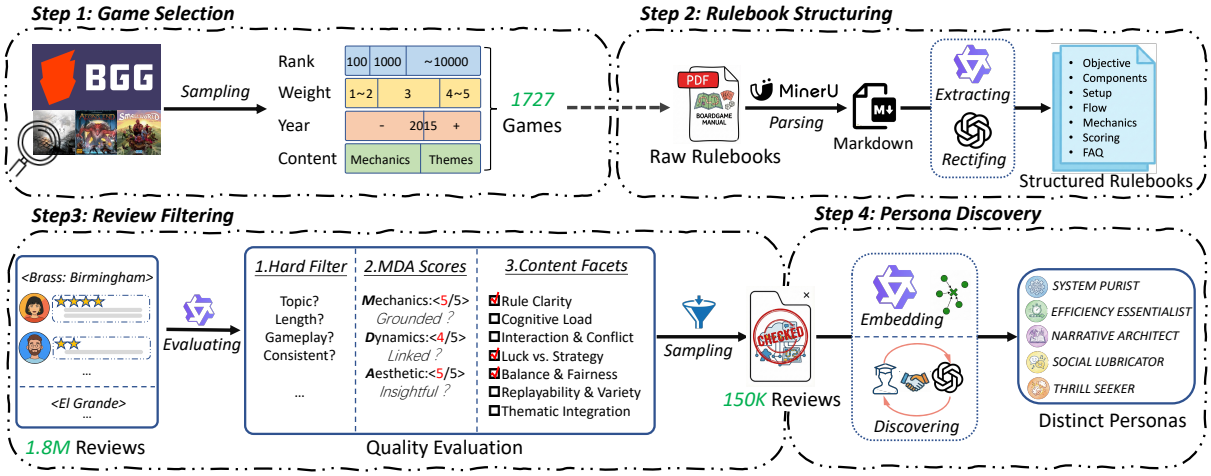


Figure 2: **The Data Construction Pipeline.** We structure 1,727 board game rulebooks and filter 1.8M reviews via multi-dimensional quality assessment, yielding 150K high-quality critiques for persona discovery.

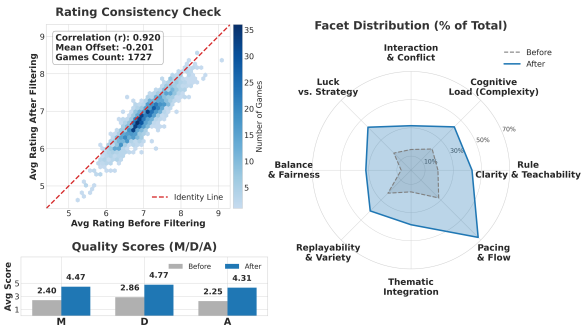


Figure 3: **Impact of the Filtering Strategy.** Our strategy enhances MDA scores and semantic coverage while preserving the original rating distribution.

formatting errors (see the rectification prompt in Appendix B.3).

3.3 Review Filtering

We aggregated a raw corpus of 1.8 million rating-comment pairs from multiple online communities (detailed in Appendix C.1). To refine this data, we employed Qwen3-235B with a multi-task prompt (detailed in Appendix C.2) to perform a comprehensive evaluation across three dimensions: (1) **Hard Filtering**, removing noise such as short texts, off-topic logistics, and rating inconsistencies; (2) **MDA Scoring**, evaluating whether specific Mechanics are linked to Dynamic interactions to derive Aesthetic experiences, and assigning quality scores across these three dimensions; and (3) **Facet Identification**, where the model mapped the content to predefined semantic topics (e.g., *Rule Clarity*, *Balance & Fairness*) to capture diverse viewpoints.

Based on these metrics, we implemented a stratified coverage-maximization strategy. We first performed stratified sampling based on original ratings to preserve sentiment fidelity (Pearson's $r = 0.920$, verified in Figure 3, Top-Left). Simultaneously, we filtered for high-quality entries (scores > 4) to significantly enhance MDA scores (Figure 3, Bottom-Left) and optimized selection to maximize semantic coverage across all facets (Figure 3, Right). This process yielded a final dataset of $\sim$150k entries (approx. 8% retention), ensuring a robust volume of 50–100 reviews per game; further details are provided in Appendix C.3.

3.4 Persona Discovery

A single "average rating" fails to capture the subjective diversity, where the rigorous complexity prized by strategists is perceived as an exhausting burden by others. To model these domain-specific cognitive attributions, we implemented a *Cluster-then-Refine* discovery pipeline, evolving raw behavioral clusters into interpretable player personas.

Discovery Pipeline. We first generated composite embeddings for all reviews using Qwen3-Embedding-8B (Zhang et al., 2025) (concatenating text with logic scores and facets; see Appendix D.1). Following K-Means clustering ($K = 15$), we employed a human-in-the-loop process where GPT-5.1 profiled representative samples, and domain experts refined these into a finalized taxonomy of five distinct personas (detailed in Appendix D.2). Using these finalized definitions, we employed GPT-5.1 to annotate the entire dataset. To ensure classification stability, we implemented a majority-vote mechanism (aggregating 3 independent inferences per review) to assign the dominant

Game	Diff.	Highest Persona (Rating)	Lowest Persona (Rating)
<*Unspeakable Words*> (Party / Word)	4.04	**Social Lubricator** (6.9) *"Party chaos!"*	**System Purist** (2.9) *"Random noise."*
<*Talisman*> (Adventure / Roll)	3.71	**Narrative Architect** (7.1) *"An epic journey."*	**System Purist** (3.4) *"Roll-and-move hell."*
<*Aeon Trespass*> (Campaign / Heavy)	2.97	**Narrative Architect** (9.0) *"Immersive masterpiece."*	**Efficiency Essentialist** (6.0) *"Feels like a job."*

Table 1: **Case Study of Aesthetic Segregation.** The rating gaps between persona groups validate that our taxonomy effectively captures distinct player preferences.

persona label. The annotation prompts are detailed in Appendix D.3.

Aesthetic Segregation. Table 1 exemplifies the distinct preference patterns captured by our taxonomy. The data reveals that party and adventure elements actively alienate *System Purists*, whereas heavy campaign games frustrate *Efficiency Essentialists* despite their thematic appeal. (See Appendix D.4 for the full Preference Matrix).

Why LLM Annotation? We attempted to train a supervised classifier (DeBERTa-v3-large (He et al., 2021)) but it proved insufficient for capturing subtle preference patterns. As analyzed in Appendix D.5, it naively misclassify a review mentioning "house rules" and "balance" as *System Purist*, failing to discern that the user is actually introducing variants to inject the high-risk, high-reward volatility typical of a *Thrill Seeker*.

4 Methodology

4.1 Problem Formulation

We formulate the task as a conditional generation problem. Given a rulebook context $\mathcal{R}$ and a target player persona $\mathcal{P}$, the objective is to generate a feedback entry $\mathcal{Y}$ (comprising a numerical rating and a textual review). Crucially, direct mapping $\mathcal{R} \rightarrow \mathcal{Y}$ ignores the *semantic gap* between static text and emergent fun. To bridge the gap between objective rules and subjective preference, we employ an *MDA-Guided Reasoning* strategy. Drawing upon the foundational game design framework (Hunicke et al., 2004), we reinterpret the MDA model originally designed to analyze gameplay loops as a causal inference chain for language models. We define a latent intermediate sequence $\mathcal{Z}$ that explicitly traces the causal path from *Mechanics* to *Dynamics*, and finally to *Aesthetics*. By decomposing the generation into $[\mathcal{R}, \mathcal{P}] \xrightarrow{\mathcal{Z}_{\text{MDA}}} \mathcal{Y}$, we force the model to simulate the runtime experience before articulating the critique, ensuring the output is logically grounded in the rules.

4.2 Synthesizing the MDA Cognitive Chain

Since the reasoning chain $\mathcal{Z}$ is latent in raw reviews, we utilize a distillation approach to recover this logic. We employed Qwen3-235B as a Teacher Model to reconstruct $\mathcal{Z}$ from high-quality review-rule pairs. The prompt (Appendix E.1) enforces a three-step cognitive flow: (1) **Step 1: Mechanics ("The What").** *"What specific content does the review explicitly mention?"* Isolate objective rule components from $\mathcal{R}$ explicitly mentioned in the review to ensure grounding. (2) **Step 2: Dynamics ("The How").** *"What Interaction or System Dynamic occurred during play?"* Infer the runtime system behaviors or interactions triggered by the mechanics identified in Step 1. (3) **Step 3: Aesthetics ("The Feel").** *"What was the final Aesthetic Experience or emotional feeling?"* Synthesize the subjective emotional outcome, modulated by the preferences of Persona $\mathcal{P}$.

Verifier-Guided Filtration Loop. We employ GPT-5.1 to judge the *entailment* between synthesized reasoning and ground-truth ratings. Guided by Appendix E.2, the verifier removes chains with sentiment contradictions or hallucinations, triggering automatic regeneration to ensure logical consistency. A sample alignment of rules, personas, and critiques appears in Appendix E.3.

4.3 Persona-Conditional Instruction Tuning

We fine-tuned the Qwen3-8B (Yang et al., 2025) backbone to maximize the joint likelihood of the MDA reasoning chain $\mathcal{Z}$ and the final critique $\mathcal{Y}$. To address the challenge of subjective heterogeneity, we do not represent $\mathcal{P}$ as a simple label. Instead, we encode the **full semantic profile** derived in Section 3.4 (including core values and interaction preferences) into the system instruction. This forces the model to treat $\mathcal{P}$ as a contextual prior that modulates the *Dynamics* $\rightarrow$ *Aesthetics* transition. Formally, we treat the concatenation $\mathcal{S} = [\mathcal{Z}; \mathcal{Y}]$ as the target sequence and minimize the standard Cross-Entropy Loss:

$$\mathcal{L} = -\sum_{t=1}^{|\mathcal{S}|} \log P(s_t \mid s_{<t}, \mathcal{R}, \mathcal{P}_{profile}). \quad (1)$$

The training was implemented using LoRA (Hu et al., 2022) on all linear layers via LLaMA-Factory (Zheng et al., 2024). Hyperparameter details are provided in Appendix E.4.

5 Experiments and Analysis

To systematically validate MeepleLM as a reliable virtual playtester, we structure our evaluation around three core research questions: (1) **RQ1 (Macro-level Alignment)** assesses whether the simulator accurately replicates community rating distributions and preference rankings. (2) **RQ2 (Micro-level Fidelity)** examines if the generated reviews maintain factual consistency with rules while exhibiting the content richness and **semantic diversity** characteristic of real players. (3) **RQ3 (Practical Utility)** investigates whether the simulated feedback provides actionable insights for design optimization and player decision support.

5.1 Experimental Setup

Dataset Splitting. We constructed a comprehensive test set of 207 games disjoint from the training corpus. To ensure representative coverage of the design landscape, we employed a stratified sampling strategy based on *BGG Weight* (Complexity 1.0–5.0) and *Average Rating* (Tier 1–5). Notably, this selection spans a wide temporal range, explicitly including 34 newly released titles (2024–2025) alongside historical classics, enabling us to assess performance on both established consensus and fresh content.

Simulation Protocols. For each game, we execute $N = 100$ simulation runs. In each run, the model takes the rulebook $\mathcal{R}$ and a specific persona $\mathcal{P}$ to generate the critique $\mathcal{Y}$. Crucially, we do not pick $\mathcal{P}$ randomly; instead, we sample personas to match the empirical proportions found in the ground-truth reviews. The specific inference prompt is provided in Appendix F.1.

Baselines. We benchmark against three state-of-the-art general LLMs (GPT-5.1, Gemini3-Pro, Qwen3-235B) and our backbone model Qwen3-8B. Detailed implementation configurations are provided in Appendix F.2.

5.2 RQ1: Macro-level Community Alignment

Evaluation Metrics. To assess whether MeepleLM aligns with the collective wisdom of the community, we employ three complementary metrics: (1) **Mean Absolute Error (MAE)** measures the absolute precision of rating predictions; (2) **Wasserstein Distance (WD)** evaluates the fidelity of the predicted score distribution against the ground truth (Villani et al., 2008); (3) **Kendall's Rank Correlation** (τ) assesses the model's ability to correctly rank games based on their perceived quality (Kendall, 1938).

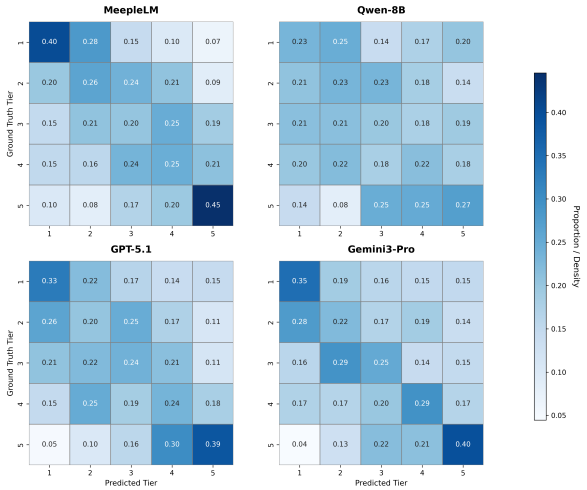


Figure 4: **Tier-wise Prediction Alignment.** MeepleLM shows a sharp diagonal concentration, effectively distinguishing quality tiers.

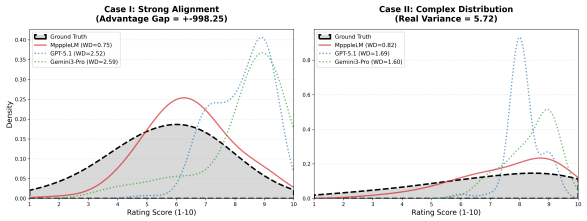


Figure 5: **Rating Density Distributions.** MeepleLM demonstrates superior distributional fidelity by accurately recovering the high variance of human consensus.

Beyond Ranking: Capturing Community Diversity. As summarized in Table 2, MeepleLM consistently achieves the best performance across all alignment metrics. This superiority is visually corroborated in **Figure 4**, where our model demonstrates a sharp diagonal concentration, effectively distinguishing high-quality outliers (Tier 1) from poor designs (Tier 5). Critically, while advanced baselines like GPT-5.1 retain some capacity for ranking ($\tau = 0.2555$), they exhibit severe *central tendency bias*, "playing it safe" by clustering predictions around the mean to minimize error. This failure to simulate authentic polarization is quantified by their high Wasserstein Distance (0.9496) compared to MeepleLM (0.2205), proving that our method captures the variance of community sentiment. In contrast, the untuned Qwen3-8B fails to establish any meaningful correlation ($\tau \approx 0$), degenerating into effectively random guessing.

Overcoming Positivity Bias. We further investigate this distributional shift in Figure 5. In sce-

Table 2: **Overall performance.** MeepleLM shows superior performance in community alignment, generation quality, and practical utility, validating the effectiveness of persona-aligned simulation for virtual playtesting.

Model	Preference Alignment (RQ1)			Review Quality (RQ2)			Utility (RQ3)
	MAE ↓	WD ↓	τ ↑	Fact. ↑	Dist-2 ↑	Div. ↑	Op-Rec ↑
GPT-5.1	0.9874	0.9496	0.2555	**99.46**	0.6934	4.26	63.44
Gemini3-Pro	1.4277	0.5092	0.2465	98.28	0.6480	3.98	57.74
Qwen3-235B	1.2288	0.6350	0.1449	98.95	0.6572	3.56	54.27
Qwen3-8B	0.8906	1.0119	0.0492	97.88	0.5936	1.58	11.39
MeepleLM(Ours)	**0.6576**	**0.2205**	**0.2817**	98.86	**0.7117**	**4.34**	**69.77**
w/o MDA	0.7395	0.4148	0.2271	91.56	0.6850	3.70	55.35
w/o Persona	0.7887	0.3630	0.1348	92.13	0.6771	3.56	53.84
w/o Rulebook	0.7043	0.5496	0.0026	59.87	0.6970	3.30	9.99

narios with polarized opinions (Case II), baselines exhibit *mode collapse*, aggressively clustering predictions around safe, high scores (7–9) and failing to capture the long-tail of negative feedback. MeepleLM, empowered by domain-specific rule understanding, successfully recovers the high variance of human consensus ($WD = 0.82$), proving its ability to represent the diverse spectrum of player sentiments rather than just a generic average.

5.3 RQ2: Content Fidelity and Diversity

Evaluation Metrics. To ensure the generated reviews are both trustworthy and rich in content, we employ three metrics covering accuracy, vocabulary, and semantic variety:

(1) **Factual Correctness (Fact.):** We implement an automated Fact-Checker using Gemini3-Flash. As detailed in Appendix F.3, the judge extracts specific claims regarding game components or mechanics from the review and strictly verifies whether they exist in the official rulebook $\mathcal{R}$. This directly measures the model's groundedness.

(2) **Lexical Diversity (Dist-2):** We use the Distinct-2 score (Li et al., 2016) to measure vocabulary richness, calculating the ratio of unique bigrams to total bigrams to detect repetitive phrasing.

(3) **Perspective Diversity (Div.):** A realistic simulator should not be a "broken record" that repeats the same opinion endlessly. To detect *semantic repetition*, we feed batches of $k = 5$ reviews (generated for the same game and persona) to a Gemini3-Flash judge(prompt details in Appendix F.4). The judge scores the batch on a 1–5 scale based on topic coverage: Low scores indicate the reviews are merely rephrasing the same point (e.g., all 5 reviews complain about "luck"); High scores indicate the reviews discuss diverse aspects such as mechanics, social interaction, and art style to mimic the natural variety of human feedback.

Results Analysis. Table 2 confirms that MeepleLM matches the factual accuracy of SOTA models while delivering superior diversity. To further illustrate this, we present the side-by-side comparison in Table 3. While Qwen3-8B defaults to a generic melodramatic tone ("theater of tears") and GPT-5.1 sounds like a detached journalist ("social WD-40"), MeepleLM authentically captures the distinct voice of each persona. By seamlessly switching from community slang (e.g., "Alpha Gamers") in social contexts to technical critique (e.g., "Variant Rules") for purists, our model proves it is not just retrieving knowledge, but truly simulating a player's perspective.

5.4 RQ3: Practical Utility

Opinion Recovery Rate (Op-Rec). To quantify the model's value as a virtual playtester, we assess its ability to forecast actual market feedback. We define Op-Rec as the recall rate of ground-truth player opinions within the simulated reviews. The evaluation employs a two-stage pipeline using Gemini3-Flash(prompts detailed in Appendix F.5): (1) **Ground Truth Mining:** The judge extracts a deduplicated set of distinct viewpoints ($\mathcal{V}_{GT}$) from historical human reviews, representing the actual "voice of the customer." (2) **Semantic Matching:** We check whether the simulated reviews generated by MeepleLM cover these specific viewpoints. As reported in Table 2, MeepleLM achieves the highest Op-Rec score, validating its utility for designers in forecasting market feedback and surfacing diverse player viewpoints.

User Study: Blind A/B Test. To validate real-world effectiveness, we conducted a controlled blind A/B test with $N = 10$ participants (demo-

Table 3: **Qualitative Case Study: Generated Reviews for <One Night Ultimate Werewolf>.** MeepleLM generates factually grounded critiques that align with specific persona sensibilities. By capturing both technical nuances and community-specific slang, our model demonstrates the semantic richness and perspective diversity.

Qwen3-8B	MeepleLM	GPT-5.1
Persona: The Social Lubricator *(Goal: Party Atmosphere, Interaction)*		
"...The table became a theater of accusations and fake tears... The mechanic was pure chaos... It's not the most elegant design, but it thrives on the energy of the group—it's a party game that rewards chaos over calculation."	*"...You really need at least 6 people though... With 4 players, the Mason role is useless. If you have alpha gamers, they will dominate the conversation... Overall, I recommend this for a quick and easy game."*	*"...As someone who treats board games as social WD-40, this is perfect. The single-night structure keeps things punchy... A fast, noisy trust-fall in social form."*
"...The night phase became a hilarious circus of whispered accusations... My friend's 'insomniac' kept checking his card... It's a perfect recipe for laughs, tension, and bonding..."	*"...Played a ton with friends... favorite game at parties. Simple enough to teach to a new player... great for people who don't know a lot about board games. The bluffing is fun..."*	*"...Basically a structured excuse for people to point fingers... That's social gold. What I love is how shifting roles (Robber, Troublemaker) give everyone a reason to speak."*
Persona: The System Purist *(Goal: Mechanics, Balance, Strategy)*		
"...One Night Ultimate Werewolf is a sharp knife with a blunt handle... The inherent randomness felt like a betrayal of the game's potential... A near-perfect match for a purist, but the luck left a bittersweet aftertaste."	*"...I like that there is no dice luck. However, the Robber and Drunk can screw up the game... I recommend adding variant rules: Drunk swaps with a Villager instead of center card."*	*"... When everyone reports honestly-then-bluffs deliberately, the deduction feels crisp. But this design lives or dies on human precision... I respect the underlying clockwork."*

graphics in Appendix G). Each evaluator reviewed 6 titles randomly selected from the test set: 3 they had previously played (*Familiar*) and 3 they had not (*Unfamiliar*). Results indicate a decisive preference for MeepleLM over GPT-5.1. In the *Familiar* scenario, our model achieved an average win rate of 78.3%, with 83.3% of participants specifically citing superior "authenticity" in capturing the emotional nuances of gameplay. In the *Unfamiliar* scenario, the average win rate remained high at 74.2%; notably, 86.7% of users preferred MeepleLM for its critical honesty, describing it as "less like marketing" and more effective for identifying potential design flaws (Detailed pairwise results and qualitative feedback are provided in Appendix G).

5.5 Ablation and Further Analysis

Ablation. To verify the contribution of each module, we evaluated three variants (detailed configurations provided in Appendix H.1). As shown in Table 2, performance drops in all cases: (1) **w/o Rulebook:** Removing rule context causes a collapse in Factual Accuracy (98.9 $\rightarrow$ 59.9), confirming that explicit grounding is non-negotiable; (2) **w/o Persona:** Replacing specific profiles with generic prompts drops ranking alignment (τ to 0.13), proving that modeling heterogeneity is essential for capturing polarized preferences; (3) **w/o MDA:** Bypassing the CoT chain results in consistently lower opinion recovery, validating that intermediate reasoning is required to bridge the gap between static text and emergent experience.

Temporal Impact. We further validated the stability by re-evaluating RQ1 on a subset excluding the 35 newly released titles (2024–2025). As detailed in Appendix H.2, performance shifts across all models are negligible, confirming that the inclusion of fresh content does not bias the assessment.

Persona Robustness. Decomposing RQ1 by player profile (Appendix H.3) highlights MeepleLM's robust performance on high-variance personas such as *The Social Lubricator* and *The Thrill Seeker*. This indicates that the framework successfully captures social dynamics and subjective "vibes" that defy pure logical deduction.

6 Conclusion

We present MeepleLM, a model that bridges the gap between static rulebooks and subjective player experiences. By curating dataset of rule-critique pairs and integrating MDA-based reasoning with data-driven player personas, we make gameplay dynamics explicit. Our experiments demonstrate that MeepleLM significantly outperforms general LLMs in capturing authentic community sentiment and actionable design insights. Ultimately, this work establishes a new paradigm for automated virtual testing of interactive systems, paving the way for experience-aware Human-AI collaboration attuned to diverse audience sensibilities.

Limitations

While MeepleLM demonstrates strong potential as a virtual playtester, we acknowledge two primary limitations that outline our roadmap for future research: (1) **Multimodal Understanding.** Currently, MeepleLM processes game rules exclusively as text. However, board games are inherently multimodal experiences where visual cues including card art, board iconography, and component design play a crucial role in immersion and usability. Future iterations will integrate visual encoders to process game assets (e.g., cards, maps, and tokens), enabling a more holistic evaluation of the game's aesthetic and functional design. (2) **Granularity of Personas.** Our current approach relies on five aggregated personas derived from community clusters, which effectively capture broad player archetypes but may overlook the unique idiosyncrasies of specific individuals. Moving forward, we aim to transition from group-level to individual-level modeling. By collecting detailed historical data from specific players, we plan to construct a granular "virtual player community," where agents can simulate the precise tastes and behaviors of real-world individuals for hyper-personalized playtesting.

Ethics Statement

Data Privacy and Usage. Our dataset is constructed from publicly available content retrieved from an online board game community (see Section 3), which is accessible to the general public. To protect user privacy, we strictly anonymize all User IDs and review identifiers, removing any personally identifiable information (PII) from the raw data. To mitigate the potential dissemination of harmful content and respect copyright considerations, we will only release the processed versions of the reviews and metadata, rather than the raw scraped content.

Human Evaluation. Our research involves collecting evaluation data from real human participants. We adhere to strict ethical guidelines to ensure their privacy, consent, and well-being. Key ethical principles include: (1) **Informed Consent**: Participants are provided with detailed information about the study's purpose, procedures, and their rights. They are explicitly informed that they can withdraw from the study at any time without any negative consequences. (2) **Data Anonymization**: To safeguard participant privacy, all collected evaluation data, including interaction logs and questionnaires, is anonymized. Personal identifiers are removed to ensure that individuals cannot be traced from the data. (3) **Data Security**: All collected data is stored securely, with access restricted to authorized research personnel only.

References

Karim Benharrak, Tim Zindulka, Florian Lehmann, Hendrik Heuer, and Daniel Buschek. 2024. Writer-defined ai personas for on-demand feedback generation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: An empirical study. *arXiv preprint arXiv:2304.00723*.

Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023. Compost: Characterizing and evaluating caricature in llm simulations. *arXiv preprint arXiv:2310.11501*.

Yoonseo Choi, Eun Jeong Kang, Seulgi Choi, Min Kyung Lee, and Juho Kim. 2025. Proxona: Supporting creators' sensemaking and ideation with llm-powered audience personas. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–32.

Yoonseo Choi, Eun Jeong Kang, Min Kyung Lee, and Juho Kim. 2023. Creator-friendly algorithms: Behaviors, challenges, and design opportunities in algorithmic platforms. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–22.

Alan Cooper. 1999. The inmates are running the asylum. In *Software-ergonomie'99: design von informationswelten*, pages 17–17. Springer.

Cong Fang, Yujie Zhu, Le Fang, Yonghao Long, Huan Lin, Yangfan Cong, and Stephen Jia Wang. 2025. Generative ai-enhanced human-ai collaborative conceptual design: A systematic literature review. *Design Studies*, 97:101300.

Jodi Forlizzi and Katja Battarbee. 2004. Understanding experience in interactive systems. In *Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques*, pages 261–268.

Mingqi Gao, Xinyu Hu, Xunjian Yin, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2025. Llm-based nlg evaluation: Current status and challenges. *Computational Linguistics*, pages 1–27.

Kaj Hansteen Izora and Christof Teuscher. 2025. Exploring the potential of large language models (llms)

to simulate social group dynamics: A case study using the board game" secret hitler". *Northeast Journal of Complex Systems (NEJCS)*, 7(2):5.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Jiale Hong, Hongqiu Wu, and Hai Zhao. 2025. Game development as human-llm interaction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4333–4354.

Chengpeng Hu, Yunlong Zhao, and Jialin Liu. 2024. Game generation via large language models. In *2024 IEEE Conference on Games (CoG)*, pages 1–4. IEEE.

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.

Robin Hunicke, Marc LeBlanc, Robert Zubek, and 1 others. 2004. Mda: A formal approach to game design and game research. In *Proceedings of the AAAI Workshop on Challenges in Game AI*, volume 4, page 1722. San Jose, CA.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Wolfgang Lehrach, Daniel Hennes, Miguel Lazaro-Gredilla, Xinghua Lou, Carter Wendelken, Zun Li, Antoine Dedieu, Jordi Grau-Moya, Marc Lanctot, Atil Iscen, and 1 others. 2025. Code world models for general game playing. *arXiv preprint arXiv:2510.04542*.

Danrui Li, Sen Zhang, Samuel S Sohn, Kaidong Hu, Muhammad Usman, and Mubbasir Kapadia. 2025. Cardiverse: Harnessing llms for novel card game prototyping. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29723–29750.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 110–119.

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, Yuxuan Lai, Chongyang Tao, and Shuai Ma. 2024. Leveraging large language models for nlg evaluation: Advances and challenges. *arXiv preprint arXiv:2401.07103*.

Wenye Lin, Jonathan Roberts, Yunhan Yang, Samuel Albanie, Zongqing Lu, and Kai Han. 2025. Gamebot: Transparent assessment of llm reasoning in games. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7656–7682.

Joakim Lindfors. 2025. Leveraging generative ai to create themed assets for games: A study on narrative and 3d asset creation with ai.

Ling Ma, Mingyao Pan, Vince Siu, Xiaoyu Chang, Jussi Holopainen, Jixing Li, and Ray LC. 2025. Follow my logic: Generative ai workflows in designing for serious table-top games. In *International Conference on Human-Computer Interaction*, pages 153–172. Springer.

Renkai Ma, Xinning Gui, and Yubo Kou. 2023. Multi-platform content creation: the configuration of creator ecology through platform prioritization, content synchronization, and audience management. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Jennifer McGinn and Nalini Kotamraju. 2008. Data-driven persona development. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1521–1524.

Junbo Niu, Zheng Liu, Zhuangcheng Gu, Bin Wang, Linke Ouyang, Zhiyuan Zhao, Tao Chu, Tianyao He, Fan Wu, Qintong Zhang, Zhenjiang Jin, Guang Liang, Rui Zhang, Wenzheng Zhang, Yuan Qu, Zhifei Ren, Yuefeng Sun, Yuanhong Zheng, Dongsheng Ma, and 42 others. 2025. Mineru2.5: A decoupled vision-language model for efficient high-resolution document parsing. *Preprint*, arXiv:2509.22186.

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*, pages 1–22.

Andrew Patrick and Md Abdullah Al Hafiz Khan. 2025. Gamegenesis: A multimodal ai revolution in board game design.

Hannah Rashkin, Elizabeth Clark, Fantine Huot, and Mirella Lapata. 2025. Help me write a story: Evaluating llms' ability to generate writing feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25827–25847.

Carles Moya Rodríguez. 2025. Opportunities in the board game market: A strategic analysis through the blue ocean theory.

Jenna Russell, Marzena Karpinska, and Mohit Iyyer. 2025. People who frequently use chatgpt for writing tasks are accurate and robust detectors of ai-generated text. *arXiv preprint arXiv:2501.15654*.

Joni Salminen, Kathleen Guan, Soon-gyo Jung, Shammur A Chowdhury, and Bernard J Jansen. 2020. A literature review of quantitative persona creation. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–14.

Joni Salminen, Bernard J Jansen, Jisun An, Haewoon Kwak, and Soon-gyo Jung. 2018. Are personas done? evaluating their usefulness in the age of digital analytics. *Persona Studies*, 4(2):47–65.

Joongi Shin, Michael A Hedderich, Bartłomiej Jakub Rey, Andrés Lucero, and Antti Oulasvirta. 2024. Understanding human-ai workflows for generating personas. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, pages 757–781.

Tsunehiko Tanaka and Edgar Simo-Serra. 2024. Grammar-based game description generation using large language models. *IEEE Transactions on Games*.

Wenjie Tang, Yuan Zhou, Erqiang Xu, Keyan Cheng, Minne Li, and Liquan Xiao. 2025. Dsgbench: A diverse strategic game benchmark for evaluating llm-based agents in complex decision-making environments. *arXiv preprint arXiv:2503.06047*.

Pittawat Taveekitworachai, Kantinan Plupattanakit, and Ruck Thawonmas. 2024. Assessing inherent biases following prompt compression of large language models for game story generation. In *2024 IEEE Conference on Games (CoG)*, pages 1–4. IEEE.

Graham Todd, Sam Earle, Muhammad Umair Nasir, Michael Cerny Green, and Julian Togelius. 2023. Level generation through large language models. In *Proceedings of the 18th International Conference on the Foundations of Digital Games*, pages 1–8.

Graham Todd, Alexander G Padula, Matthew Stephenson, Éric Piette, Dennis J Soemers, and Julian Togelius. 2024. Gavel: Generating games via evolution and language models. *Advances in Neural Information Processing Systems*, 37:110723–110745.

Cédric Villani and 1 others. 2008. *Optimal transport: old and new*, volume 338. Springer.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Dingyi Yang and Qin Jin. 2025. What matters in evaluating book-length stories? a systematic study of long story evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16375–16398.

Qing Ru Yong and Alex Mitchell. 2023. From playing the story to gaming the system: Repeat experiences of a large language model-based interactive story. In *International Conference on Interactive Digital Storytelling*, pages 395–409. Springer.

Pengfei Yu, Dongming Shen, Silin Meng, Jaewon Lee, Weisu Yin, Andrea Yaoyun Cui, Zhenlin Xu, Yi Zhu, Xingjian Shi, Mu Li, and 1 others. 2025. Rpgbench: Evaluating large language models as role-playing game engines. *arXiv preprint arXiv:2502.00595*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, and 1 others. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Mingzhe Zheng, Dingjie Song, Guanyu Zhou, Jun You, Jiahao Zhan, Xuran Ma, Xinyuan Song, Ser-Nam Lim, Qifeng Chen, and Harry Yang. 2025. Cmlbench: A framework for evaluating and enhancing llm-powered movie scripts generation. *arXiv preprint arXiv:2510.06231*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

A Dataset Statistics Details

In this section, we provide a comprehensive statistical breakdown of the 1,717 board games selected for our dataset. These statistics validate that our sampling strategy successfully captured a diverse range of difficulty levels, quality standards, historical eras, and gameplay mechanisms.

A.1 Distribution Analysis

Figure 6 presents the distributions of Complexity, Rating, Publication Year, and BGG Rank.

- **Complexity (Weight):** As shown in Figure 6(a), the complexity distribution is nearly normal (Skewness: 0.29) with a mean of 2.57. We intentionally preserved a balanced spectrum: *Light* games (Weight 1–2) account for 24.5%, while *Heavy/Very Heavy* games (Weight > 3) comprise 28.5%. This ensures the model learns to adapt its critique depth to the cognitive load of the game.

- **Rating:** Figure 6(b) shows the rating distribution (Mean: 7.22, StdDev: 0.51). The distribution is slightly left-skewed, focusing on games generally considered "playable" to "excellent" (66% are rated > 7.0). This filtering removes low-quality noise while retaining enough variance for comparative analysis.

- **Publication Year:** Illustrated in Figure 6(c), the dataset reflects the modern board game renaissance, with a median publication year of 2013. The coverage spans from classic designs (pre-2000, $N = 159$) to contemporary hits (2015+, $N = 755$), specifically including 34 cutting-edge titles released in 2024–2025.

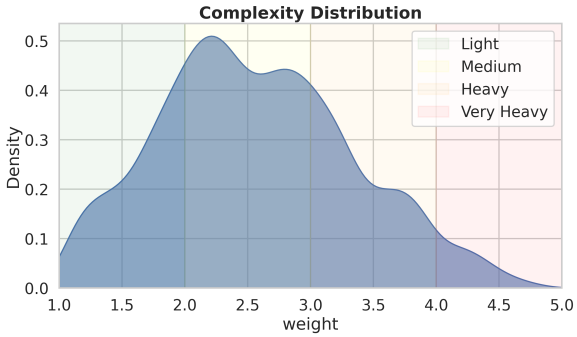
- **Rank Coverage:** Figure 6(d) highlights the market representativeness. While we include 83% of the Top 100 "elite" games to ensure high-quality training data, over 50% of the dataset ($N = 909$) consists of "long-tail" games (Rank > 1000), preventing the model from overfitting to universally acclaimed masterpieces.

A.2 Content Diversity: Mechanics and Themes

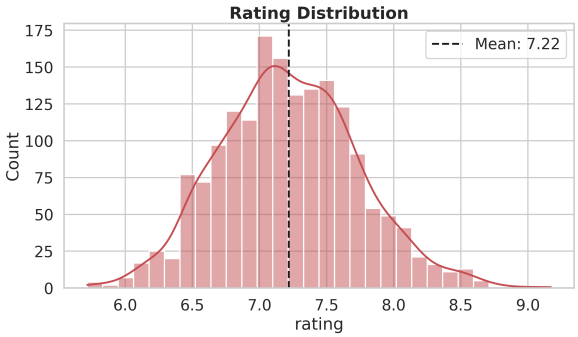
To ensure the model can generate grounded reviews for various gameplay styles, we analyzed the mechanics and themes tags. Figure 7 visualizes the prevalent categories.

- **Mechanics:** The dataset features **192 unique mechanics** with an average of 6.35 mechanics per game, indicating high systemic depth. As shown in Figure 7(a), the most frequent mechanics include *Hand Management* (38.4%), *Dice Rolling* (29.0%), and *Variable Player Powers* (26.2%), which are foundational to modern game design.

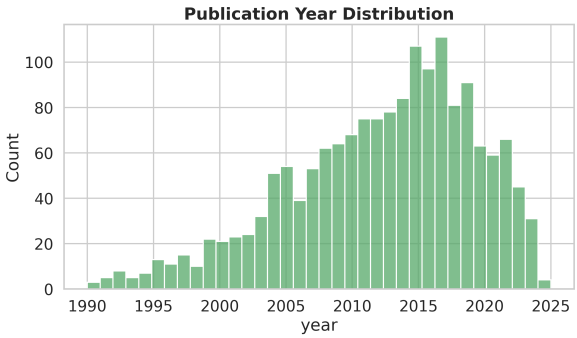
- **Themes:** We identified **81 unique themes**. Figure 7(b) shows a blend of abstract strategy themes (e.g., *Economic*, 20.9%) and immersive narrative themes (e.g., *Fantasy*, 18.8%; *Sci-Fi*, 12.0%). This diversity requires the model to contextually adapt its vocabulary (e.g., discussing "profits" vs. "damage").



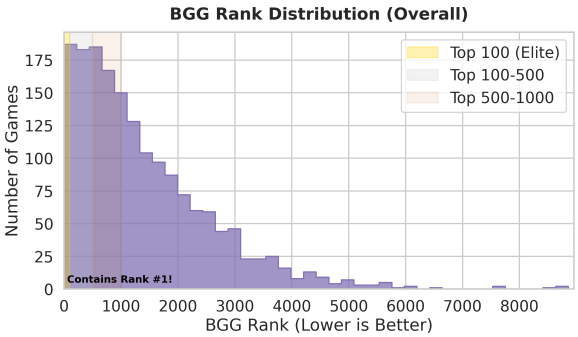
(a) Complexity (Weight) Distribution



(b) Average Rating Distribution

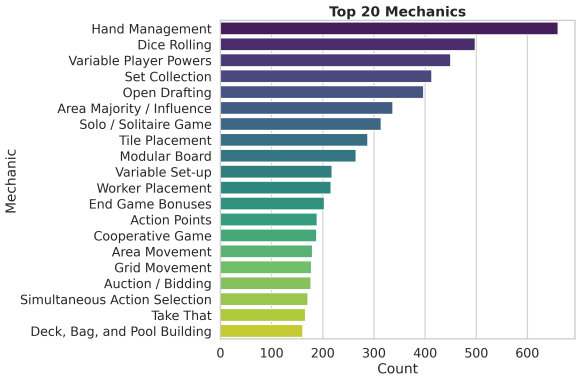


(c) Publication Year Distribution

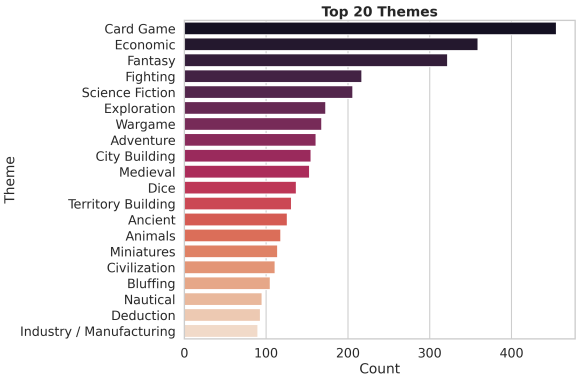


(d) BGG Rank Distribution

Figure 6: **Distributions of Key Metadata Attributes.** for the 1,717 selected games. The dataset covers a wide spectrum of difficulty (a), focuses on decent-to-excellent quality games (b), emphasizes modern board game designs (c), and spans both elite and long-tail rankings (d).



(a) Top 20 Mechanics



(b) Top 20 Themes

(c) Mechanics Word Cloud

(d) Themes Word Cloud

Figure 7: **Analysis of Game Content.** (a) and (b) display the top 10 mechanics and themes, demonstrating that the dataset covers the fundamental building blocks of modern board games. (c) and (d) provide a holistic view of the terminological diversity present in the corpus.

B Rulebook Structuring Details

To convert the raw Markdown rulebooks (converted by Mineru) into structured knowledge, we employed Qwen-3. We utilized a specific prompt to ensure the model extracts information strictly from the source text without hallucination, organizing it into a standardized Markdown format.

B.1 Extraction Prompt

Figure 8 displays the system prompt used. The prompt enforces strict constraints to use *only* existing information from the uploaded file.

B.2 Structured Rulebook Example

Figure 9 demonstrates a sample of the structured output. This standardized text serves as the knowledge base for the review generation model.

B.3 Rectification Prompt

Figure 10 presents the rectification prompt used by GPT-5.1. This stage acts as a verification layer, cross-referencing the structured draft against the source text to correct hallucinations or omissions.

Prompt: Rulebook Structuring

```
System Instruction:
You are an expert board game rules analyst. Your task is to reorganize the provided raw
    Markdown content into a structured format.

Constraints (Strict Grounding):
1. Source Only: You must ONLY use information present in the input text. Do NOT add any
    external knowledge or hallucinate rules.
2. Format: Output the result in clean Markdown format using the specific headers defined below.

3. Completeness: If a section is not mentioned in the text, write "Not Mentioned" under that
    header.

Required Output Structure:
## 1. Lore & Objective
(Extract the thematic setting and victory conditions)
## 2. Components
(List physical assets like cards, boards, tokens)
## 3. Setup
(Step-by-step initial configuration)
## 4. Gameplay Flow
(Turn structure, phases, and round sequence)
## 5. Core Mechanics
(Key interaction rules, player actions, restrictions)
## 6. Scoring & End Game
(How points are calculated and how the game ends)
## 7. FAQ or Edge Cases
(Specific clarifications found in the text)

Input Text:
{RAW_MARKDOWN_CONTENT}
```

Figure 8: **System Prompt for Structuring Raw Rulebooks.** It enforces a standard Markdown schema and strict grounding to the source text.

Structured Rulebook Example <Clank!: A Deck-Building Adventure>

```
### 1. Lore & Objective
Players are rival thieves attempting to infiltrate Dragon Keep, steal valuable Artifacts, and
     escape alive. The dragon Nicki guards her hoard —fiercelynoise (Clank!) and stolen
     treasures increase her rage. The ultimate goal is to become the Greatest Thief in the
     Realm by escaping with an Artifact and the most Victory Points (VP).

**Primary Objectives:**
- Steal at least **one Artifact—**failure to do so results in automatic loss.
- Accumulate the highest VP total from Artifacts, Gold, acquired tokens, and deck cards.
- Survive above the Grass Line (depths boundary) or escape the dungeon to score points.

**Winning Conditions:**
- Players knocked out **below the Grass Line** score 0.
- Players knocked out **above the Grass Line** score their collected loot (if they have an
     Artifact).
- Escaping players receive a **20 VP Mastery Token**.
- Game ends immediately when the **Countdown Track reaches the Skull** or all players have
     exited/knocked out.
- Ties broken by highest-value Artifact; further ties unresolved.

---

### 2. Components
- **Double-sided Game Board** (Front: recommended for first game)
- **Dungeon Deck** (100 cards)
- **Reserve Cards**:
  - 15 Mercenary
  - 15 Explore
  - 12 Secret Tome
  - 1 Goblin (always available)
- **Player Decks**: 4 starting decks (10 cards each):
  - 6 Burgle
  - 2 Stumble
  - 1 Sidestep
  - 1 Scramble
- **Tokens & Markers**:
  - 120 Clank! Cubes (30 per player color)
  - 24 Black Dragon Cubes
  - Dragon Bag
  - Dragon Marker
  - Mastery Tokens (4)
- **Treasures & Secrets**:
  - 7 Artifacts —(530 VP)
  - 11 Major Secrets
  - 18 Minor Secrets
- **Market Items**:
  - 2 Master Keys (5 VP)
  - 2 Backpacks (5 VP)
  - 3 Crowns (10, 9, 8 VP)
- **Other**:
  - 3 Monkey Idols
  - Gold Tokens (1 and 5)
  - Player Pawns (4)
  - Rage Track (on board)
  - Countdown Track (top of board)
  - Health Meter (bottom of board)

---

### 3. Setup
1. **Player Setup**:
   - Each player takes:
     - 30 Clank! cubes (personal supply)
     - Matching pawn
     - 10-card starting deck
```

```
   - Shuffle deck, draw 5 cards.
   - Place pawn on top-left entrance space.

2. **Board Setup**:
   - **Artifacts**: Place 7 face up on numbered spaces.
     - 3 players: Shuffle and remove 1 randomly.
     - 2 players: Remove 2 randomly.
   - **Major Secrets**: Shuffle face down, place 1 on each marked space. Return extras to box.
   - **Minor Secrets**: Shuffle face down, place 2 stacked on each marked space.
   - **Market Items**:
     - 2 Master Keys (stacked)
     - 2 Backpacks (stacked)
     - 3 Crowns (10, 9, 8 in descending order)
   - **Monkey Idols**: Place 1 in each space in Monkey Shrine Room.
   - **Mastery Tokens**: Place 1 per player at top-left flag.
   - **Gold**: Place in bank near board.
   - **Dragon Marker**:
     - 4 players: Start on first space
     - 3 players: Second space
     - 2 players: Third space
   - **Dragon Bag**: Place 24 black cubes inside.
   - **Reserve Stacks**: Place near board:
     - 1 Goblin
     - 15 Mercenary
     - 15 Explore
     - 12 Secret Tome
   - **Dungeon Row**: Shuffle Dungeon Deck, deal 6 face-up cards.
     - Replace any with Dragon Attack symbol; shuffle replaced cards back.
   - **Starting Clank!**:
     - 1st player: 3 cubes in Clank! Area
     - 2nd: 2 cubes
     - 3rd: 1 cube
     - 4th: 0 cubes

3. **First Player**: Roll die or choose randomly.

---

### 4. Gameplay Flow
Each turn follows this sequence:

1. **Play All Cards**: Play all 5 cards in hand, in any order.
2. **Take Actions** (in any order, multiple times):
   - Acquire Cards (Skill)
   - Use Devices (Skill)
   - Fight Monsters (Swords)
   - Buy Market Items (Gold)
   - Move (Boots)
   - Gain Gold / Clank! / Secrets / Artifact
3. **End of Turn Phase**:
   - Discard all played cards.
   - Draw 5 new cards (shuffle discard if needed).
   - Refill Dungeon Row to 6 cards.
   - **If refill reveals Dragon Attack symbol**:
     - Dragon attacks once (even if multiple symbols).
     - Move all Clank! cubes to Dragon Bag.
     - Draw cubes = 'Dragons Rage Track value (+1 per Danger symbol).
     - Black cubes: set aside. Colored cubes: place on matching 'players Health Meter.
4. **Dragon Rage Triggers**:
   - +1 space when:
     - Artifact is taken
     - Dragon Egg (Minor Secret) is revealed

5. **Exit or Knockout Triggers Countdown**:
   - First player to exit dungeon or be knocked out moves to Countdown Track.
   - On their next turn, they move forward and trigger effects:
     - Spaces −24: Instant Dragon Attack with +1, +2, +3 extra draws
     - Space 5: All remaining players in dungeon are instantly knocked out
```

6. **Game End**: Triggered when:
   - Countdown Track reaches Skull
   - All players have exited or been knocked out

---

5. Core Mechanics

Resources
- **Skill**: Acquire cards from Dungeon Row or Reserve.
- **Swords**: Fight monsters (Dungeon Row or tunnels).
- **Boots**: Move 1 tunnel per Boot.

Movement Rules
- **Normal Tunnel**: 1 Boot
- **Footprint Icon**: 2 Boots
- **Monster Icon**: Pay 1 Sword per icon, or take 1 damage per icon
- **Lock Icon**: Requires Master Key
- **Arrow Tunnel**: One-way only
- **Crystal Cave**: Upon entry, **cannot use Boots** for rest of turn (teleporting allowed)
- **Wrap-around Tunnels**: Connect opposite edges; cost 1 Boot

Actions
- **Acquire Card** (Blue/Yellow Banner):
  - Pay Skill →place in discard pile (joins deck)
- **Use Device** (Purple Banner):
  - Pay Skill →use effect immediately →discard to Dungeon discard pile
- **Fight Monster** (Red Banner):
  - Pay Swords →gain reward →discard to Dungeon discard pile
  - **Goblin (Reserve)**: Can be fought multiple times; not discarded
- **Buy Market Item**:
  - Cost: 7 Gold each
  - Items: Crown (VP), Backpack (carry +1 Artifact), Master Key (unlock tunnels)
  - Unlimited purchases per turn

Clank!
- Cards or actions may add Clank! cubes to central Clank! Area.
- Negative Clank! removes cubes from area or offsets future gains.
- Unused negative Clank! is lost at turn end.

Dragon Attacks
- Triggered **only** when refilling Dungeon Row reveals **at least one Dragon symbol**.
- All Clank! cubes go into Dragon Bag.
- Draw cubes = 'Dragons current Rage Track value + number of Danger symbols in Dungeon Row.
- Colored cubes = damage to that player.
- Exited/knocked-out players are immune.

Health & Knockout
- Damage places cubes on Health Meter (left to right).
- **Fully filled meter** = knocked out.
  - **Below Grass Line**: Lose (0 points)
  - **Above Grass Line**: Score loot if holding Artifact
- Healing: Use Potions or effects to return cubes to supply.

Artifacts & Carrying
- Can only carry **1 Artifact** unless Backpack is owned (1 extra per Backpack).
- **Cannot drop or switch** Artifact once picked up.
- Taking Artifact:
  - Move Dragon +1 on Rage Track
  - Triggers Dragon Egg if Minor Secret is revealed

---

6. Scoring & End Game
Scoring Includes:
- Artifact VP
- All acquired tokens (Crown, Chalice, Backpack, etc.)

```
- Gold: 1 VP per Gold
- Cards in deck: VP in top-right corner
- **Mastery Token (20 VP)**: Only if exited dungeon via own movement/teleport

**Excluded from Score**:
- Players knocked out **below Grass Line** →0 points
- Exited/knocked-out players: No further turns or card effects

**Tiebreaker**:
1. Highest-value Artifact
2. (Unspecified if still tied)

---

### 7. FAQ or Edge Cases
- **Can you drop an Artifact?** No. Once picked up, 'its yours until game end.
- **Can you buy multiple Market items per turn?** Yes, any number, even same type.
- **Can you take multiple tokens per room?** Only one per entry. Must exit and re-enter.
- **Does Artifact in Market cost Gold?** No. Only Crown, Backpack, Key cost Gold.
- **When does Dragon attack?** Only when **new** Dragon symbol is revealed during **Dungeon
    Row refill**. Preexisting symbols do not re-trigger.
- **Crystal Cave movement?** Cannot use Boots after entry, even via teleport.
- **Running out of Clank! cubes?** Cannot choose to take damage in tunnels. Cannot be forced
    to add Clank! (free pass until healed).
- **Healing during turn?** Yes, via Potions or effects.
- **Order of card play?** Effects resolve in real time; order does not block conditional
    bonuses (e.g., Swagger gains Skill for all Clank! made that turn).

---
```

Figure 9: **Example of Structured Rulebook Data.** This content spans multiple pages, preserving the full details extracted from the original PDF.

Prompt: Rulebook Rectification (Expert-in-the-Loop)

```
System Instruction:
You are a meticulous Board Game Rule Editor. Your task is to verify and rectify a "Draft
    Rulebook" against the original "Source Content".

Inputs:
1. Source Content: Raw text parsed from the official rulebook PDF.
2. Draft Rulebook: A structured Markdown version generated by an automated parser.

Rectification Tasks:
1. Accuracy Check: Compare specific numbers, card counts, and setup instructions. If the Draft
     says "deal 5 cards" but the Source says "deal 4", CORRECT it.
2. Completeness: Ensure no critical sections (especially "End Game Triggers" or "Tie-Breakers
    ") are missing from the Draft.
3. Logical Consistency: Fix any contradictions introduced during the structuring process.
4. Formatting: Ensure the output strictly follows the standardized Markdown hierarchy.

Constraints:
- Output ONLY the fully rectified Rulebook in Markdown.
- Do NOT add conversational text (e.g., "Here is the fixed version").
- Do NOT invent rules not present in the Source Content.

Input Data:
[SOURCE CONTENT]:
{RAW_SOURCE_TEXT}

[DRAFT RULEBOOK]:
{QWEN_GENERATED_STRUCT}
```

Figure 10: **Rectification Prompt Used by GPT-5.1.** It requires the model to cross-reference the generated draft against the source text to ensure numerical accuracy and logical completeness.

C Review Processing Details

C.1 Data Sources

To ensure diverse perspective coverage, we aggregated raw user reviews from multiple online communities through professional data outsourcing services. The sources encompass prominent digital tabletop platforms such as Board Game Arena[2] and Tabletopia[3], alongside specialized enthusiast forums like GStone[4] and QPBG[5]. Given the heterogeneous scoring systems across these sites, we normalized all collected ratings to a standardized 1.0–10.0 scale.

C.2 Quality Annotation Prompt

To implement the "Design-Logic Quality Scoring" described in Section 3.3, we used the prompt shown in Figure 13. It enforces a strict evaluation criterion based on the utility of the review for game designers.

C.3 Statistical Validation

We analyzed the statistical properties of the 150K retained reviews to ensure they serve as an unbiased yet information-dense proxy for the original population.

Distributional Fidelity. As shown in Figure 11, the filtered dataset maintains a high degree of alignment with the original ratings (Pearson $r = 0.92$, Spearman $\rho = 0.91$). This confirms that our filtering strategy preserves the global consensus on game quality. Notably, we observed a slight negative mean shift (-0.20), which suggests the successful removal of "low-effort hype" (e.g., empty 10/10 ratings), resulting in a more critical and objective set.

Information Density Word count analysis reveals a "Polarization Ratio" of 1.24x: reviews at the rating extremes (1 & 10) contain significantly more text (avg. 195.7 words) compared to mid-range reviews (avg. 158.2 words). This indicates that the dataset prioritizes strong signals—users provide the most detailed structural feedback when they are passionately engaged, ensuring the model learns clear causal links for both design flaws and successes.

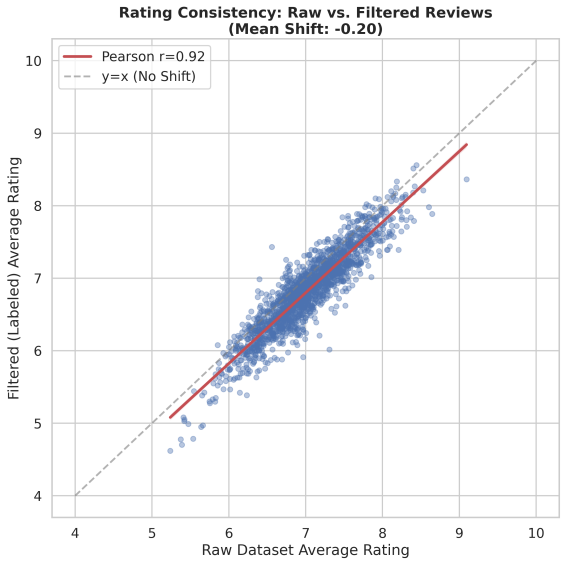


Figure 11: **Rating Correlation (Original vs. Filtered).**

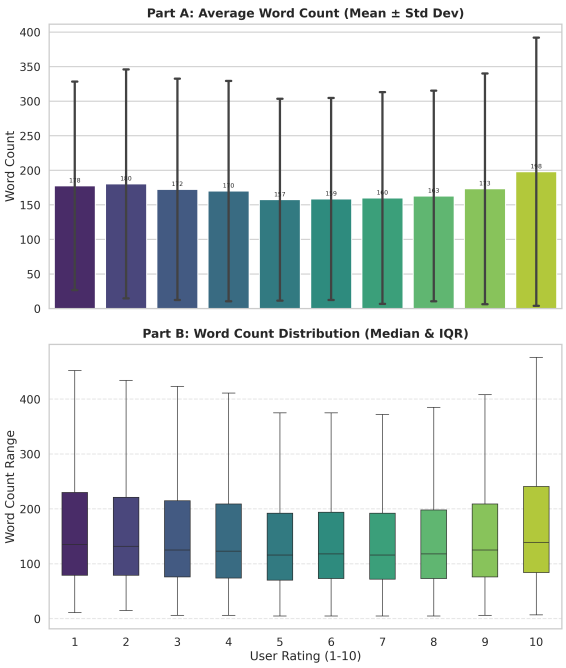


Figure 12: **Word Count Statistics by Rating.**

[2] https://en.boardgamearena.com
[3] https://tabletopia.com
[4] https://www.gstonegames.com
[5] https://qpbg.com

Prompt: Review Quality Scoring

```
You are an expert Board Game Research Assistant constructing a high-quality reasoning dataset.
For each review (given a numeric rating and review text), you must output a JSON object
    strictly following the definitions below.

### 0. CRITICAL SCORING RULES (DECOUPLING)
**Treat each score INDEPENDENTLY.** Do not allow high scores in one dimension to bleed into
    others (No "Halo Effect").
- **High Anchoring (5) does NOT imply High Attribution**: A user might list rule names like a
    manual (Anchoring: 5) but explain nothing about the experience (Attribution: 1).
- **High Attribution (5) does NOT imply High Anchoring**: A user might deeply explain the
    cause of their fun (Attribution: 5) using only vague terms like "the pieces" (Anchoring:
    2).
### 1. TASK DEFINITIONS
#### A. Hard Filters (Boolean)
Determine if the review should be discarded. Set "is_valid": false if ANY of the following
    apply:
- **Irrelevant**: Discusses ONLY shipping, damaged boxes, Kickstarter delivery, or customer
    service.
- **Visuals Only**: Discusses ONLY artwork, miniatures, or card stock quality without
    mentioning gameplay mechanics or experience.
- **Too Short**: Contains fewer than 20 words (insufficient logic).
- **Rating Mismatch**: The sentiment of the text drastically contradicts the numeric rating (e.
    g., a glowing review with a rating of 1, or a hateful rant with a rating of 10).
#### B. Quality Scores (Use the FULL 1-5 Scale)
You must use the full range of integers (1, 2, 3, 4, 5). Do not default to just 1/3/5.
1. **mechanism_anchoring** (Specificity)
   - 1 (Vague): "Fun game", "Good strategy". No specific terms.
   - **2 (Basic):** Mentions basic components like "cards", "board", "points" but no mechanism
       names.
   - 3 (Generic): Mentions standard mechanics e.g., "worker placement", "deck building".
   - **4 (Detailed):** Describes specific game flow or unique twists but misses the exact
       rulebook terminology.
   - 5 (Precise): Cites exact rule names/components e.g., "The Tekhenu obelisk wheel", "The
       Cult Track".
2. **causal_attribution** (Reasoning - CORE METRIC)
   - 1 (No Logic): "I hated it." / "Best game ever." (Pure emotion).
   - **2 (Implied):** "It's too long and boring." (Reason implies cause, but vague).
   - 3 (Simple Link): "I didn't like it *because* the downtime was too long." (Direct X->Y).
   - **4 (Strong Logic):** "The downtime is caused by the analysis paralysis in the auction
       phase." (Identifies mechanism -> issue).
   - 5 (Deep Systemic): "Because turn order resets based on travelers, long-term planning
       feels impossible, leading to a chaotic experience." (Mechanism -> Dynamic -> Aesthetic)
       .
3. **constructiveness** (Utility)
   - 1 (Useless): Empty complaints or blind praise.
   - **2 (Valid Complaint):** "The combat feels unfair." (Identifies a problem area, but
       subjective).
   - 3 (Actionable): "The endgame drags on too long." (Specific pain point).
   - **4 (Analytical):** "The blue faction is strong because of their starting resource." (
       Analyzes *why* it's a pain point).
   - 5 (Insightful/Solution): Offers a fix or deep balance critique: "They should cap rounds
       to 8 to fix the pacing."
#### C. Content Facets (Multi-label List)
Identify which aspects are primarily discussed. Ignore background mentions.
Choose ONLY from:
- "Rule Clarity & Teachability": rules readability, ambiguity, teaching difficulty.
- "Cognitive Load (Complexity)": weight, brain-burn, analysis paralysis.
- "Interaction & Conflict": take-that, blocking, table talk, multi-player solitaire.
- "Luck vs. Strategy": randomness, dice/card luck, mitigation.
- "Balance & Fairness": start player advantage, runaway leader, faction balance.
- "Replayability & Variety": setup variability, scripted play.
- "Thematic Integration": Ludonarrative harmony (mechanics fit theme). Exclude pure art praise.

- "Pacing & Flow": downtime, game length feel (drag/tight), end triggers.
### 2. OUTPUT SCHEMA
```

```
Return a JSON object with this EXACT structure:
{
  "is_valid": boolean,
  "filter_reason": string or null, // If is_valid is false, state the reason (e.g., "Too Short
      ", "Irrelevant"). If true, null.
  "scores": {
    "mechanism_anchoring": int(1-5),
    "causal_attribution": int(1-5),
    "constructiveness": int(1-5)
  },
  "facets": [string, string, ...] // List of matching facets
}
### 3. FEW-SHOT EXAMPLES
Input:
Rating: 6
Comment: "It involves moving pawns and collecting cards, but the randomness in drawing cards
    makes it feel like I have no control."
Output:
{
  "is_valid": true,
  "filter_reason": null,
  "scores": {
    "mechanism_anchoring": 2, // "pawns", "cards" (Basic terms) -> 2
    "causal_attribution": 4, // Randomness -> No control (Strong Logic) -> 4
    "constructiveness": 2 // Valid complaint but subjective -> 2
  },
  "facets": ["Luck vs. Strategy"]
}
Input:
Rating: 8
Comment: "Excellent worker placement game. The tight economy forces you to plan ahead."
Output:
{
  "is_valid": true,
  "filter_reason": null,
  "scores": {
    "mechanism_anchoring": 3, // "worker placement" (Generic) -> 3
    "causal_attribution": 3, // tight economy -> plan ahead (Simple Link) -> 3
    "constructiveness": 1 // Praise only -> 1
  },
  "facets": ["Cognitive Load (Complexity)"]
}
Input:
Rating: 4
Comment: "The 'Corruption' mechanic punishes you too hard. It creates a death spiral where if
    you mess up turn 1, you sit there for 2 hours. It feels broken."
Output:
{
  "is_valid": true,
  "filter_reason": null,
  "scores": {
    "mechanism_anchoring": 5, // "Corruption" (Specific Rule) -> 5
    "causal_attribution": 5, // Mechanic -> Dynamic (Death spiral) -> Feeling -> 5
    "constructiveness": 4 // Strong analysis of the problem -> 4
  },
  "facets": ["Balance & Fairness", "Pacing & Flow"]
}
You will be given a list of {batch_size} board game reviews.
You MUST generate a JSON ARRAY containing EXACTLY {batch_size} JSON objects, one for each
    review.
Do not add more or fewer objects than the number of reviews provided.
REVIEWS (JSON list):
{reviews_json}
```

Figure 13: **Annotation Prompt for Scoring Review Quality.** It mirrors the observation-analysis-iteration loop.

D Persona Discovery Details

D.1 Feature Construction & Clustering

To ensure the clustering algorithm captures the cognitive depth of the reviewer rather than just keyword overlap, we pre-processed each review into a composite text string before embedding.

Composite Input Logic. We injected the quantitative metrics derived in Section 3.3 directly into the text representation. Each review was formatted using the following structured template:

```
[SENTIMENT: {Tier}] [FOCUS: {Facets}] :: {Raw Review Content}
```

The metadata fields were populated based on the following rules:

- **Sentiment Tier:** Discretized based on the normalized rating R: labeled as *"Positive"* ($R \geq 8$), *"Negative"* ($R \leq 4$), or *"Neutral"* ($5 \leq R \leq 7$). This guides the embedding to group reviews by satisfaction level.
- **Focus Facets:** A comma-separated list of dimensions derived directly from the facet-scoring model detailed in Section 3.3.

Clustering Result. Figure 14 visualizes the T-SNE projection of these embeddings. The inclusion of explicit **Sentiment** and **Focus** tags helped clearly separate reviewers based on their fundamental evaluation criteria and satisfaction thresholds, effectively mitigating the ambiguity of surface-level keywords.

D.2 Persona Descriptions and Statistics

Figure 15 shows the distribution and average rating for each group. Figure 16 provides the detailed definitions for each persona used in the annotation process, outlining their core motivations and specific mechanical preferences.

D.3 Discovery Prompts

We employed a two-phase prompting strategy to translate raw numerical clusters into an annotated dataset.

Phase 1: Persona Profiling. After obtaining the 15 initial clusters, we sampled the top-20 central reviews from each cluster. As shown in Figure 17, we prompted **GPT-5.1** to analyze these samples and summarize the distinct "Player Persona" they

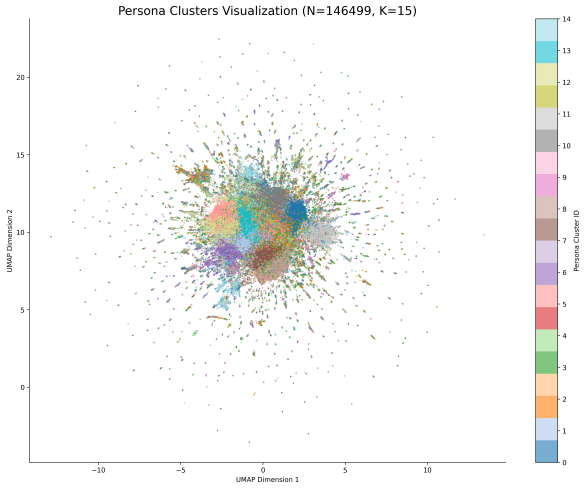


Figure 14: **Visualization of Composite Embeddings.** Colors indicate the 15 initial clusters, which were later merged into the 5 final personas by domain experts.

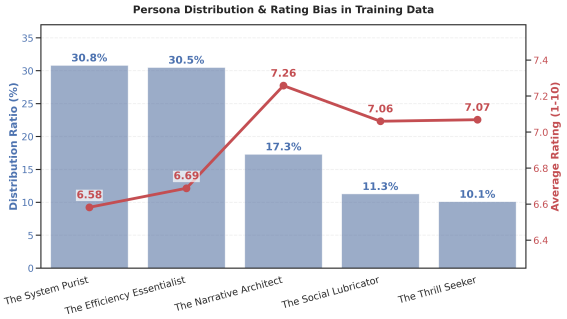


Figure 15: **Distribution of Personas in the Dataset.** *System Purists* (Avg 6.58) are the harshest critics, while *Narrative Architects* (Avg 7.28) are the most generous.

represent. Domain experts analyzed the semantic coherence of each cluster, synthesizing overlapping groups to establish the final five persona definitions.

Phase 2: Dataset Labeling. Once the 5 distinct personas were finalized (as defined in Appendix D.2), we needed to propagate these labels to the entire dataset. Figure 18 displays the classification prompt used with **GPT-5.1** to annotate each review.

D.4 Preference Matrix (Extended Case Study)

To validate the distinctiveness of each persona, we performed a frequency analysis of mechanics in their highest-rated vs. lowest-rated games. Figure 19 and Figure 20 detail the "Lift" metric, highlighting mechanics that are disproportionately favored or disliked by specific groups.

D.5 Semantic Ambiguity Analysis

To demonstrate the limitations of standard supervised classifiers, we present a detailed case study from our error analysis. As noted in Section 3.4, a DeBERTa-v3-large classifier trained on cluster seeds achieved only ∼50% accuracy.

Table 4 illustrates the core issue: **Keyword vs. Intent Misalignment**. Standard models over-index on technical vocabulary (e.g., *"balance"*, *"rules"*), failing to detect when users repurpose these terms to describe their exact opposites—such as using "house rules" to introduce **high-stakes volatility** rather than mechanical fairness.

Review Input (Raw Text)
"A **sentimental** rating of course. We combined parts of the first Lord of the Rings game with this one, working on making the deck more *balanced*. We had numerous *house rules*, that in my opinion, made it such a great gaming experience. For example, when attacking, **triple sixes** gave you an **extra kill**. Double sixes on defense, you GAINED an army, lose NONE. To even that out though, if you roll double ones, you LOSE 3. Remember, this is all before I learned about board games."
Standard Classifier Prediction: System Purist ✗ **Reasoning:** The model is misled by surface-level design keywords such as *"balanced"* and *"house rules"*. It incorrectly infers that the user is focused on game balance, mechanical rigor, or improving the system's logic.
Ground Truth / MeepleLM Label: Thrill Seeker ✓ **Reasoning:** The review explicitly frames the experience as *"sentimental"* and prioritizes dramatic, high-variance moments (*"triple sixes"*, *"extra kill"*, *"LOSE 3"*). The user's "house rules" were not created to fix the system's logic, but to inject more chaos and excitement into the gameplay, which is the defining trait of a Thrill Seeker.

Table 4: **Case Study of Semantic Ambiguity.** The standard classifier fails by latching onto mechanical keywords ("balanced"), while the LLM correctly identifies the underlying motivation for excitement ("triple sixes", "sentimental").

Detailed Persona Definitions & Preferences

```
"The System Purist": """
    * **Core Motivation:** Intellectual superiority & Control. They want to prove they can
        beat the system through pure logic.
    * **Profile:** Loves heavy/crunchy decisions. Zero tolerance for luck (hates dice output
        randomness). Obsessed with balance (hates first-player advantage).
    * **Interaction:** Likes indirect competition (blocking), hates chaotic direct conflict (
        take-that).
    * **Keywords:** "Optimization", "No luck", "Perfect information", "Tight", "Punishing".
""",
"The Efficiency Essentialist": """
    * **Core Motivation:** Maximize ROI (Fun/Time). Seeks "Flow".
    * **Profile:** Hates "Fiddliness" (setup, shuffling, bookkeeping). Values elegance (simple
         rules, deep strategy). Pragmatic about rules (will house-rule to fix pacing).
    * **Interaction:** Fast-paced. Hates Downtime (Analysis Paralysis).
    * **Keywords:** "Elegance", "Streamlined", "Downtime", "Fiddly", "Smooth".
""",
"The Narrative Architect": """
    * **Core Motivation:** Immersion & Epic Experience. Mechanics serve the theme.
    * **Profile:** Loves growth (leveling up, empire building, tech trees). Wants 4X/RPG feels
         but within reasonable time.
    * **Interaction:** Cooperative or thematic negotiation/trade. Not calculating pure math.
    * **Keywords:** "Theme", "Immersion", "Story", "Epic", "Journey", "Flavor".
""",
"The Social Lubricator": """
    * **Core Motivation:** Human Connection & Emotional Resonance. Game is an excuse to
        socialize.
    * **Profile:** Needs low barrier to entry (accessible to non-gamers). Hates "Alpha Gamers"
         (quarterbacking). Prioritizes experience over scoring.
    * **Interaction:** High social interaction (bluffing, laughter, party games).
    * **Keywords:** "Party game", "Laughs", "Interaction", "Easy to teach", "Group dynamic".
""",
"The Thrill Seeker": """
    * **Core Motivation:** Dopamine & Emotional Rollercoaster.
    * **Profile:** Embraces risk (Push-your-luck). Needs fast pacing (if I lose, let me
        restart instantly). Active agency in gambling.
    * **Interaction:** Schadenfreude (enjoying opponents busting) and epic comebacks.
    * **Keywords:** "Push your luck", "Excitement", "Tension", "Gambling", "High stakes".
"""
```

Figure 16: **Behavioral Profiles of the Five Discovered Personas.** Each profile defines core motivations and specific likes/dislikes regarding game mechanisms.

Prompt : Persona Profiling (GPT-5.1)

```
You are a User Researcher specialized in board games. I will provide you with a set of game
    reviews that belong to the same Player Cluster. Your task is to analyze these reviews and
     synthesize a Persona Profile.
Based ONLY on the reviews above, define this player persona using the following JSON schema.
    Be specific and refer to the evidence in the text.
{
  "persona_name": "Creative Name (e.g., The Dice Hater, The Euro Optimizer)",
  "core_motivation": "Why do they play? (e.g., Mathematical efficiency)",
  "preferred_mechanics": "What do they like/dislike? (e.g., Loves worker placement, hates dice
      )",
  "interaction_style": "How do they interact? (e.g., Dislikes direct conflict)",
  "deal_breakers": "What makes them rate a game low? (e.g., Bad rulebooks, too much luck)",
  "system_prompt_description": "A concise, first-person description for an LLM to roleplay
      this user. (e.g., 'You are a hardcore strategy gamer who values low luck and high skill
      ceiling. You are critical of imbalance...')"
}
```

Figure 17: **Profiling Prompt for Interpreting Cluster-Central Samples.** This qualitative analysis guided the definition of the final 5 personas.

Prompt: Persona Labeling (GPT-5.1)

```
You are an expert User Researcher specialized in board games. Your task is to classify board
    game players into EXACTLY ONE of the following 5 distinct personas.

### VALID PERSONAS (Choose ONE)
{PERSONA_DEFINITIONS}

### TASK
Analyze the review. Assign the review to one of the above personas.

### OUTPUT FORMAT
Return a JSON OBJECT. Key: "LLM_persona_name". Value: Must be one of the 5 bolded titles above.

Example: {"LLM_persona_name": "The System Purist"}

Process these {batch_size} reviews. Return a JSON ARRAY of {batch_size} objects.

REVIEWS:
{reviews_json}
```

Figure 18: **Labeling Prompt for Annotating the Full Dataset.** It maps each review to one of the 5 finalized personas based on the review's content and sentiment.

1. THE SYSTEM PURIST

LOVED Games

Examples: Guards of Atlantis II, 1817, Pax Renaissance: 2nd Edition, Age of Innovation...

Key Mechanics:

- **Hand Management** (Freq: 33%, Lift: **1.5x**)
- **End Game Bonuses** (Freq: 20%, Lift: **4.6x**)
- **Auction / Bidding** (Freq: 17%, Lift: **3.0x**)
- **Area Movement** (Freq: 13%, Lift: **2.1x**)
- **Market** (Freq: 13%, Lift: **13.5x**)

HATED Games

Examples: The Werewolves of Miller's Hollow, The Mind, Dungeon Run, Wavelength...

Key Mechanics:

- **Dice Rolling** (Freq: 47%, Lift: **1.6x**)
- **Hand Management** (Freq: 33%, Lift: **1.5x**)
- **Cooperative Game** (Freq: 23%, Lift: **1.8x**)
- **Role Playing** (Freq: 13%, Lift: **10.9x**)
- **Player Elimination** (Freq: 10%, Lift: **5.0x**)

2. THE EFFICIENCY ESSENTIALIST

LOVED Games

Examples: Aeon's End: The New Age, Maria, The Crew: Mission Deep Sea, 7 Wonders (Second Edition)...

Key Mechanics:

- **Hand Management** (Freq: 43%, Lift: **1.9x**)
- **Area Majority / Influence** (Freq: 27%, Lift: **2.2x**)
- **Dice Rolling** (Freq: 20%, Lift: **0.7x**)
- **Campaign / Battle Card Driven** (Freq: 17%, Lift: **4.6x**)
- **Cooperative Game** (Freq: 13%, Lift: **1.0x**)

HATED Games

Examples: WarCraft: The Board Game, The Elder Scrolls V: Skyrim – The Adventure Game, Android, Car Wars...

Key Mechanics:

- **Dice Rolling** (Freq: 47%, Lift: **1.6x**)
- **Hand Management** (Freq: 37%, Lift: **1.6x**)
- **Cooperative Game** (Freq: 17%, Lift: **1.3x**)
- **Action Points** (Freq: 13%, Lift: **1.5x**)
- **Modular Board** (Freq: 10%, Lift: **1.9x**)

3. THE NARRATIVE ARCHITECT

LOVED Games

Examples: Aeon Trespass: Odyssey, Skull, BattleCON: Devastation of Indines, Aeon's End: The New Age...

Key Mechanics:

- **Cooperative Game** (Freq: 37%, Lift: **2.9x**)
- **Hand Management** (Freq: 30%, Lift: **1.3x**)
- **Dice Rolling** (Freq: 23%, Lift: **0.8x**)
- **Auction / Bidding** (Freq: 17%, Lift: **3.0x**)
- **Area Majority / Influence** (Freq: 17%, Lift: **1.4x**)

HATED Games

Examples: The Great Split, Blackbeard: The Golden Age of Piracy, Zombies!!!, Aquädukt...

Key Mechanics:

- **Hand Management** (Freq: 37%, Lift: **1.6x**)
- **Dice Rolling** (Freq: 27%, Lift: **0.9x**)
- **Simultaneous Action Selection** (Freq: 17%, Lift: **6.1x**)
- **Communication Limits** (Freq: 13%, Lift: **7.8x**)
- **Race** (Freq: 13%, Lift: **10.4x**)

Figure 19: **Detailed Mechanism Preferences (Part 1).** Analysis of *System Purist*, *Efficiency Essentialist*, and *Narrative Architect*.

4. THE SOCIAL LUBRICATOR

LOVED Games

Examples: American Rails, Napoléon: The Waterloo Campaign, 1815, Egizia, Arydia: The Paths We Dare Tread...

Key Mechanics:

- **Dice Rolling** (Freq: 40%, Lift: **1.4x**)
- **Area Majority / Influence** (Freq: 30%, Lift: **2.5x**)
- **End Game Bonuses** (Freq: 17%, Lift: **3.8x**)
- **Cooperative Game** (Freq: 17%, Lift: **1.3x**)
- **Area Movement** (Freq: 13%, Lift: **2.1x**)

HATED Games

Examples: Feudum, Clippers, Fürstenfeld, World of Warcraft: The Adventure Game...

Key Mechanics:

- **Hand Management** (Freq: 37%, Lift: **1.6x**)
- **Action Points** (Freq: 23%, Lift: **2.7x**)
- **Area Majority / Influence** (Freq: 17%, Lift: **1.4x**)
- **Dice Rolling** (Freq: 17%, Lift: **0.6x**)
- **Network and Route Building** (Freq: 13%, Lift: **4.5x**)

5. THE THRILL SEEKER

LOVED Games

Examples: 20th Century, Advanced Squad Leader, Gaia Project, Empires in Arms...

Key Mechanics:

- **Dice Rolling** (Freq: 23%, Lift: **0.8x**)
- **End Game Bonuses** (Freq: 17%, Lift: **3.8x**)
- **Hexagon Grid** (Freq: 17%, Lift: **2.2x**)
- **Area Movement** (Freq: 17%, Lift: **2.6x**)
- **Contracts** (Freq: 13%, Lift: **4.1x**)

HATED Games

Examples: Lanterns: The Harvest Festival, Through the Ages: A New Story of Civilization, Dixit, Abalone...

Key Mechanics:

- **Hand Management** (Freq: 23%, Lift: **1.0x**)
- **Action Points** (Freq: 13%, Lift: **1.5x**)
- **Auction: Dutch** (Freq: 10%, Lift: **36.5x**)
- **Dice Rolling** (Freq: 10%, Lift: **0.3x**)
- **Open Drafting** (Freq: 10%, Lift: **1.1x**)

Figure 20: **Detailed Mechanism Preferences (Part 2).** Analysis of *Social Lubricator* and *Thrill Seeker*. The high "Lift" values (e.g., 62.4x for Prisoner's Dilemma) indicate strong predictive power of these features for persona identification.

E Cognitive Simulation (CoT) Details

This section provides the implementation details for constructing the Chain-of-Thought training data.

E.1 CoT Construction Prompt

To convert raw reviews into structured reasoning chains, we fed the Rulebook and the Raw Review into Qwen-3-Instruct using the prompt displayed in Figure 22.

E.2 CoT Verifier Prompt

To rigorously filter out hallucinatory or logically incoherent training data, we employed a Verifier-Guided Filtration strategy. Figure 21 presents the exact system instruction used for the **Verifier Model (GPT-5.1)**. Acting as a "Senior Logic Auditor," the model is tasked with strictly evaluating the causal entailment between the synthesized MDA reasoning (specifically the Aesthetic derivation) and the ground-truth rating, rejecting any chain where the logic contradicts the user's numerical score or hallucinates rules absent from the source text.

E.3 CoT Data Example

Figure 23 demonstrates a processed training instance. During instruction tuning, the model inputs the Rules and Persona, and learns to sequentially generate the '<thought>' block (The MDA Chain) followed by the '<review>' block.

E.4 Hyperparameter Configuration

To facilitate the reproducibility of our experiments, Table 5 provides the detailed hyperparameter configuration used for the Persona-Conditional Instruction Tuning phase. The model was fine-tuned using the **LLaMA-Factory** framework (Zheng et al., 2024) . We enabled the "Slow Thinking" mechanism, which incorporates the generated Chain-of-Thought tokens into the loss calculation, ensuring the model optimizes the reasoning process alongside the final output.

Hyperparameter	**Value**
Model & Environment	
Backbone Model	Qwen-3-8B
Framework	LLaMA-Factory
Context Window	16,384 tokens
Attention Mechanism	Flash Attention v2
LoRA Configuration	
Target Modules	All Linear Layers
LoRA Rank (r)	32
LoRA Alpha (α)	64
LoRA Dropout	0.1
Optimization	
Learning Rate	5.0×10^{-5}
LR Scheduler	Cosine
Warmup Ratio	0.03
Optimizer	AdamW
Num Epochs	3
Batching & Strategy	
Per-Device Batch Size	2
Gradient Accumulation	8
Effective Global Batch Size	128
Reasoning Mode	Slow Thinking
Dataset Template	qwen

Table 5: **Training Hyperparameters for Persona-CoT.**

Prompt: Logical Consistency Verification

```
You are a Senior Logic Auditor for a Board Game Research Database.
Your goal is to detect **Hallucinations** and **Logical Inconsistencies** in synthetic data.

### INPUT DATA
- User Review: {review_text}
- Ground Truth Rating: {rating} / 10
- Synthesized MDA Chain: {generated_json}

### VERIFICATION CRITERIA (Pass/Reject)
Assess the "Synthesized MDA Chain" against the following strict rules. You must be critical.

1. **Grounding Check (The "What")**
   - Does "Step 1: content_extraction" only contain facts explicitly present in the User
       Review?
   - [CRITICAL]: Reject if it cites game mechanics/rules that are NOT mentioned in the user's
       text (Hallucination).

2. **Causal Logic Check (The "How")**
   - Does "Step 2: dynamic_interaction" logically follow from the mechanics identified?

3. **Sentiment Alignment (The "Feel")**
   - Does "Step 3: experience_outcome" logically support the **Ground Truth Rating**?
   - [FAILURE MODE A]: The chain describes the experience as "frustrating" or "broken", but
       the Rating is High (>7). -> REJECT.
   - [FAILURE MODE B]: The chain describes "thrilling tension", but the Rating is Low (<4). ->
        REJECT.

### DECISION LOGIC
- If ANY of the above checks fail, the status is "REJECT".
- If the reasoning is sound, grounded, and matches the score, the status is "PASS".

### OUTPUT SCHEMA
Return a single JSON object:
{
  "status": "PASS" | "REJECT",
  "reason": "Brief explanation of the error (e.g., 'Sentiment Mismatch: Reasoning describes
      chaos as negative, but rating is 9/10').",
  "suggestion": "Brief hint for regeneration (optional)."
}
```

Figure 21: **Prompt for the Consistency Verifier.** The model audits synthesized reasoning chains to ensure they are factually grounded in the review text and causally aligned with the ground-truth rating.

Prompt: MDA Cognitive Extraction

```
You are an expert Ludologist (Game Researcher) analyzing board game reviews.
Your task is to perform a **"Reverse Experience Reconstruction"** based on a user's review and
     the game's rules.

### ANALYSIS TASK (Chain of Thought)
If the rules are valid, reconstruct the player's experience strictly following the **"What ->
    How -> Feel"** flow.
You must output a JSON object containing a `thought_chain` with exactly these three steps:

**Step 1: content_extraction (The "What")**
* What specific content does the review explicitly mention?
* Identify the **Theme**, **Mechanics**, or **Specific Details** (e.g., "The combat cards", "
    The trading phase", "The zombie theme") referenced by the user.
* *Constraint:* Do not guess. Only cite what is in the text.

**Step 2: dynamic_interaction (The "How")**


* Based on the rules and the user's description, what **Interaction** or **System Dynamic**
    occurred during play?
* How did the mechanics listed in Step 1 actually function? (e.g., Did it cause downtime? Did
    it create a tense standoff? Did it force players to lie to each other?)

**Step 3: experience_outcome (The "Feel")**
* What was the final **Aesthetic Experience** or emotional feeling?
* Why did the dynamic in Step 2 result in a positive or negative judgment?
* *Context:* Use the provided **Player Persona** to explain *why* they reacted this way (e.g.,
     "As a System Purist, they hated this randomness," or "As a Social Lubricator, they loved
     this chaos").

### OUTPUT SCHEMA
{
  "thought_chain": {
    "content_extraction": "...",
    "dynamic_interaction": "...",
    "experience_outcome": "..."
  }


}

### CONTEXT
**Target Game Rules (Excerpt):**
{rule_content}

**Player Persona (Reference Only):**
{persona_def}

**User Review (Ground Truth):**
"{review_text}"

### TASK
Perform the Reverse Experience Reconstruction.
Analyze the review to generate the "thought_chain" (content_extraction -> dynamic_interaction
    -> experience_outcome).
```

Figure 22: **Instruction Prompt for Extracting Latent MDA Reasoning.** The Teacher Model uses this prompt to extract the latent MDA reasoning chain from raw reviews.

Data Sample

```
[INPUT]

Game Rule: "In El Grande, players are Spanish nobles (Grandes) vying for influence . . . "

Persona: "The System Purist . . . "

Comment: "Game in a Nutshell: Basic area-majority game that gets spiced up with a power card
     type auction for turn order and action cards that allow special events on every players
     turn. Played as: 4,5 player Played with: Base game only. Pros: Great execution in game
     design were simple rules with a light splash of variability, and good player interaction
     make a game that hasn't been topped in decades. The action cards have a wide breadth of
     special actions and pull off some very juicy decisions while never feeling contrived.
     Cons: The pizzazz is all in the game-play... that is to say, we're not winning any beauty
      contests. While the rules are intuitive enough you would think anyone could play, the
     interaction is probably a bit much for most non-gamers (and a good chunk of gamers).
     Lastly, and not really an issue of the game, but \"boooo\" to the publisher for
     continually printing this game in a \"big box\" format. The expansions are generally


     regarded as junk, but here we are wasting materials and money to put them in every copy
     of the box. I don't appreciate the \"up-charge\" when all I want to play is the base game.
      Other: The game is great with 4 players but best with 5. I have no interest in playing
     it with 3. I truly believe this is one of the best games ever designed. The reason I don'
     t rate it higher is a matter of personal taste in that I prefer a little more freedom
     than is offered in the limited action card selection. That said, I wouldn't ask El Grande
      to change... it's perfect just the way it is. Best Use: Still the gold standard for area-
     majority games."

[OUTPUT]

content_extraction: "The review should highlight the game's core mechanics: area-majority
     scoring, Power Cards determining turn order (described as a 'power card type auction'),
     and Action Cards that enable special actions during placement. The user highlights the
     interaction between these systems, particularly how Action Cards create 'juicy decisions'
      without feeling contrived. They also reference player count dynamics (preference for 5
     players, avoidance of 3), criticize the game's physical components (big box, lack of
     aesthetic appeal), and express a personal desire for more freedom in action selection."



dynamic_interaction: "The Power Card auction creates a tightly balanced, deterministic turn
     order system where players bid with unique numbers −(113) over time, eliminating luck and
      ensuring perfect information. This directly feeds into the strategic depth: players must
      optimize not only their current turn's card play but also manage long-term sequencing of
      Power Cards. The Action Cards introduce variability through curated special abilities (e.
     g., moving the Grande, early scoring), which interact with area control by enabling
     tactical disruptions or positioning. In 5-player games, the full use of all Power Card
     values and Action Card stacks maximizes system interplay, increasing indirect competition
      through precise blocking and timing. The Castillo and Secret Disks add layers of hidden
     commitment and bluffing without randomness, aligning with a control-focused playstyle."

experience_outcome: "As a player motivated by intellectual superiority and control, the
     reviewer deeply appreciates the absence of luck and the 'games reliance on pure strategic
      optimization. The deterministic Power Card auction and tight action economy satisfy their
      need for perfect information and balanced competition. While they personally wish for
     broader action freedom, they recognize that the constraints are essential to the 'games
     precision and —elegancehence calling it 'perfect just the way it is.' Their critique of
     the box size and aesthetics reflects a purist disdain for unnecessary commercialization,
     reinforcing their identity as someone who values systemic purity over presentation. The


     high praise despite minor personal preferences underscores that the game delivers the
     intense, logical, and punishingly fair experience they crave.
```

Figure 23: **Sample of a Generated Reasoning Chain.** A sample derived from a review of *El Grande*. The generated reasoning chain correctly identifies the reviewer as an **analytical veteran who reveres elegant, deterministic game design and functional purity**.

F Experimental Setup Details

F.1 Simulation Inference Prompt

Figure 24 presents the complete system instruction used during the inference stage to generate persona-conditioned feedback ($\mathcal{Y}$).

This prompt aggregates the target persona profile ($\mathcal{P}$), the rulebook context ($\mathcal{R}$), and strict behavioral guidelines. Crucially, the "Simulation Guidelines" section is designed to mitigate stereotypical behavior by explicitly encouraging nuance, such as allowing for "guilty pleasures" or acknowledging diverse tastes within a single persona group, thereby enhancing the ecological validity of the generated critiques.

F.2 Model Deployment

All experiments were conducted with a consistent temperature setting of $T = 0.7$ to ensure comparable generation diversity. The specific deployment configurations are as follows:

- **Local Deployment:** `Qwen3-8B` and `Qwen3-235B-A22B-Instruct-2507` were deployed locally using the `vLLM` inference framework.
- **API Access:** `GPT-5.1-high` and `Gemini-3-pro-high` were accessed via their respective official APIs.

F.3 Factual Correctness Judge

Figure 25 illustrates the instruction for the **Rule Hallucination Detector**. We utilized **Gemini-3-Flash** for this task due to its long-context capability, allowing it to ingest the full rulebook $\mathcal{R}$ to verify specific claims in the generated review.

The judge classifies each extracted factual claim into specific categories based on the evidence found in the rulebook. To compute the final **Rule Accuracy** metric, we aggregate these labels as follows:

- **Correct Claims:** We consider a claim valid if it is labeled as `SUPPORTED` (explicitly found in the text) or `INFERRED` (a correct logical summary of the mechanics).
- **Hallucinations:** Claims labeled as `CONTRADICTED` (conflicting with rules).

Accordingly, the final accuracy score is calculated as the ratio of validated claims to the total number of extracted claims:

$$\text{Rule Accuracy} = \frac{N_{\texttt{SUPPORTED}} + N_{\texttt{INFERRED}}}{N_{\text{Total Claims}}} \quad (2)$$

Empirically, we observed that the generated reviews are rich in mechanical detail, with the judge typically extracting between **10 to 20 checkable claims per review**. This high density of factual assertions ensures that the accuracy score reflects a comprehensive audit of the generated content, rather than a check on a trivial or sparse summary.

F.4 Perspective Diversity Judge

Figure 26 presents the instruction for the **Perspective Diversity Judge**. This metric penalizes "Echo Chamber" behavior (where the model repeats the same point endlessly) and rewards broad coverage of diverse gameplay dimensions (e.g., mechanics, social interactions, and theme), ensuring the simulated persona reflects the multifaceted nature of a real player.

F.5 Opinion Recovery Evaluation

To verify whether MeepleLM captures specific, actionable feedback relevant to game designers (RQ3), we established a two-stage evaluation pipeline using **Gemini-3-Flash**.

Pipeline & Metric Calculation. The process consists of two steps:

1. **Ground Truth Mining:** First, we employ the LLM as a qualitative analyst to extract a set of distinct, non-redundant viewpoints ($\mathcal{V}_{GT}$) from the real human reviews in the test set. The prompt for this step is shown in Figure 27.
2. **Semantic Matching:** Next, we use a semantic match evaluator to determine which viewpoints in $\mathcal{V}_{GT}$ are successfully covered by the model's generated reviews. The prompt is presented in Figure 28.

The final **Opinion Recovery Rate (Op-Rec)** is calculated as the ratio of unique viewpoints successfully recalled by the simulation:

$$\text{Op-Rec} = \frac{|\mathcal{V}_{\text{matched}}|}{|\mathcal{V}_{GT}|} \times 100\% \quad (3)$$

where $\mathcal{V}_{\text{matched}}$ represents the subset of ground-truth viewpoints that were identified as semantically present in the generated output.

Prompt: Persona-Conditional Simulation

```
[SYSTEM PROMPT]
You are an expert Board Game Player Simulation Engine.
Current Active Persona: **{target_persona}**
**Your Goal:** Post a **comment** and a rating for the game.
You are NOT writing a formal review article. You are just sharing your quick thoughts after a
     game night.
**PERSONA PROFILE (General Tendency):** {p_def}
**SIMULATION GUIDELINES (CRITICAL):**
1. **Persona is a Bias, Not a Straitjacket:**
   - This persona represents your *general* gaming preferences, but real players are complex.
       Do not act like a one-dimensional caricature.
   - It is possible for a player to have **"Guilty Pleasures"** (e.g., enjoying a game that
       goes against their usual type) or **"Unexpected Disappointments"** (e.g., disliking a
       game that perfectly fits their profile).
2. **Embrace Diversity:**
   - Within the "{target_persona}" group, there is a wide spectrum of opinions.
   - Some players are **purists** (rejecting anything outside their genre), while others are
       **omnivorous** (appreciating good design regardless of genre).
   - You have the freedom to simulate any point on this spectrum.
3. **Ground the Review in Dynamics & Authentic Feeling:**
   - Do not just list mechanics; describe the **interactions** they created at the table (e.g.,
        "The voting mechanic caused a hilarious shouting match" vs "There is a voting mechanic
       ").
   - Connect these dynamics to your **emotional response**. Did the game feel tense?
       Frustrating? Triumphant?
   - Your rating should reflect this specific **experiential quality**, balancing your
       personal taste with the game's ability to deliver a memorable moment.
**REQUIRED OUTPUT FORMAT:**
You must output ONLY a single valid JSON object.
JSON Schema:
{
  "persona": "{target_persona}",
  "rating": Integer (1-10),
  "review": "String (A realistic review. It does not always need to be negative if the genre
      doesn't match, nor always positive if it does. Simulate a genuine reaction.)"}
[USER MESSAGE]
**Task:** Read the Game Rules below.
**Action:** Simulate a realistic review for this game from the perspective of **{
     target_persona}**.
**Game Rules:**{Rulebook Text}
Rules analysis complete. Now, simulate the review:
1. **Determine Your Stance:** As **{target_persona}**, how does this specific game land for
     you?
   - Is it a **"Guilty Pleasure"**? (e.g., "I usually hate party games, but this mechanic made
        me laugh.")
   - Is it a **"Respectful Pass"**? (e.g., "Great design, just not for me. ")
   - Is it a **"Perfect Match"** or a **"Design Failure"**?
2. **Write the Review:**
   - Focus on the **dynamics** (interactions at the table) and **emotions** (tension, joy,
       frustration).
   - Avoid generic stereotypes. Write like a real person with complex tastes.
3. **Output:** Output ONLY the valid JSON object.
**Required Output Template:**
```json
{
  "persona": "{target_persona}",
  "rating": [Integer],
  "review": "[Your review text...]"}
CONSTRAINTS: Length: Target 150-200 words, but significant variance (20-400 words) is
     mandatory to reflect real human diversity.
```

Figure 24: **Full Inference Prompt Structure.** The model receives a System Message defining the persona and guidelines, followed by a User Message containing the specific game rules and the final trigger instructions to ensure formatting compliance.
```

Prompt: Factual Correctness Judge

```
[SYSTEM MESSAGE]
You are a strict Board Game Fact-Checker.
Your SOLE task is to verify if the **Components, Mechanics, and Rules** mentioned by the
    player explicitly exist in the official **Rulebook**.
**1. SCOPE (What to Check):**
- **Existence:** Did the player mention a specific component (e.g., "Dice", "Meeple")? Check
    if it is in the component list.
- **Mechanism:** Did the player mention a specific rule (e.g., "Drafting", "Auction")? Check
    if this mechanism exists.
- **Procedure:** Did the player describe a flow (e.g., "Deal 3 cards")? Check if the number/
    action is correct.
**2. EXCLUSION CRITERIA (CRITICAL - IGNORE ALL OF THESE):**
- **IGNORE FEELINGS & OPINIONS:** Do not check statements like "It feels tense", "It is
    balanced", "It is fun", "The luck is annoying".
- **IGNORE STRATEGIC ADVICE:** Do not check "Always buy red cards first".
- **IGNORE EXTERNAL COMPARISONS:** Do not check "Like Catan".
- **IGNORE NARRATIVE ERRORS:** Do not check "I thought X but I was wrong".
**3. VERIFICATION LOGIC:**
Extract ONLY factual claims about *what the game is* or *how it plays*. Ignore *how it feels*.
* **SUPPORTED (Factually Correct):**
    - The mentioned component/rule explicitly exists in the text.
    - *Example:* "There are 5 distinct factions." -> Rules lists 5 factions. -> **SUPPORTED**
    - *Example:* "You roll dice to attack." -> Rules mention dice in combat. -> **SUPPORTED**
* **INFERRED (Logical Summary of Rules):**
    - The player summarizes a rule mechanism correctly without quoting it verbatim.
    - *Example:* "This is a hidden role game." -> Rules describe 'Traitors' and 'Secret
        Agendas'. -> **INFERRED** (Correct classification of mechanics).
    - *Note:* Do NOT use this for feelings. "This is a tense game" -> **IGNORE** (Not a rule
        summary).
* **CONTRADICTED (Hallucination/Factual Error):**
    - **Non-existent Entity:** Mentions a component that is NOT in the game.
        - *Example:* "I rolled the dice." -> Rulebook has NO dice. -> **CONTRADICTED**
    - **Wrong Number/Action:** Describes a rule incorrectly.
        - *Example:* "You verify specific claims." -> Rules say "Verify ALL claims". -> **
            CONTRADICTED**
**Output Format (JSON List):**
[
  {{ "claim": "Player mentions using an 'Auction' mechanic", "status": "SUPPORTED", "reason":
      "Rulebook Section 3 describes the 'Bidding Phase'." }},
  {{ "claim": "Player mentions 'Dice'", "status": "CONTRADICTED", "reason": "Component list
      only includes Cards and Tokens. No dice found." }}
]
[USER MESSAGE]
**Official Rulebook Context:**
{rulebook_text}
**Player Review:**
{review_text}
**TASK:**
1. **STRIP AWAY** all adjectives, emotions, and opinions (e.g., ignore "brutal", "fun", "
    random feel").
2. **FOCUS** only on the Nouns (Components) and Verbs (Actions/Rules).
3. **VERIFY**: Do these things exist in the Rulebook?
   - If user says "The dice combat is bad" -> Only check if "Dice Combat" exists. Ignore "bad".

   - If user says "There are no dice" -> Check if there are truly no dice.

Output ONLY the JSON list.
```

Figure 25: **System Prompt for Factual Verification.** The judge strictly compares mechanical claims in the review against the ground-truth rulebook, ignoring subjective opinions.

Prompt: Perspective Diversity Judge

```
[SYSTEM MESSAGE]
You are a Lead Game Designer analyzing playtest feedback.
You are given {len(reviews_batch)} reviews for the **SAME GAME** written by the **SAME PERSONA
    **.

**OBJECTIVE:**
Determine if the AI model is capable of **True Perspective Diversity** or if it suffers from
    **Semantic Repetition**.
We need to distinguish between "surface-level variation" (changing words) and "deep structural
     shifts" .

**FRAMEWORK: (Mechanics, Dynamics, Aesthetics)**
1. **Mechanics:** Rules, components, math.
2. **Dynamics:** Run-time behavior, player interaction, pacing.
3. **Aesthetics:** Emotional response, theme, sensory experience.

**STRICT SCORING CRITERIA (1-5):**

* **1 (Echo Chamber / Mode Collapse):**
   * The reviews are effectively clones. They cite the exact same rules and express the exact
        same sentiment.
   * *Example:* All 5 reviews complain about the "dice rolling combat".

* **2 (Surface Rephrasing):**
   * The core topic is identical, but the wording is different.
   * *Example:* Review A: "The combat is random." Review B: "Fighting relies too much on luck
       ." (Same point).

* **3 (Intra-Layer Variation):**
   * The reviews discuss **different features**, but they stay within the **SAME MDA layer**.
   * *Example:* Review A talks about *Dice* (Mechanic). Review B talks about *Cards* (
       Mechanic).
   * *Verdict:* Good, but lacks depth/breadth.

* **4 (Cross-Layer Shifts):**
   * The reviews shift focus across **DIFFERENT layers**.
   * *Example:* Review A analyzes the *Auction math* (Mechanics). Review B discusses the *
       Table Talk/Bluffing* (Dynamics).
   * *Verdict:* High diversity.

* **5 (Panoramic / Holistic):**
   * **Rare and Exceptional.** The set covers Mechanics, Dynamics, AND Aesthetics
       distinctively.
   * It feels like the persona is looking at the game through a —kaleidoscopeeach review
       reveals a completely new dimension (e.g., Logic vs. Emotion vs. Social).

**Output Format (JSON):**
{{
   "score": ,
   "reason": " "
}}

[USER MESSAGE]
**Game ID:** {game_id}
**Persona:** {persona}
**Generated Samples (Batch of {len(reviews_batch)}):**
{reviews_text_block}
**Task:**
Rate the Perspective Diversity (1-5). Be **STRICT**.
Output ONLY the JSON.
```

Figure 26: **System Prompt for Perspective Diversity Scoring.** The judge evaluates a batch of reviews to determine if the model exhibits semantic collapse (repeating the same points) or true diversity (shifting focus across MDA layers).

Stage 1: Viewpoint Mining Prompt

```
[SYSTEM MESSAGE]
You are a Qualitative Data Analyst specializing in Board Games.
Your task is to maintain and expand a **Comprehensive List of Distinct Viewpoints** regarding
    a specific game, based on player reviews.

**GOAL:**
Read the **New Reviews** and extract any *NEW* arguments, mechanics mentions, or specific
    experiences that are NOT already covered in the **Current Viewpoints List**.
Merge them into the list.

**RULES:**
1. **Be Extensive:** If a review mentions a specific detail (e.g., "The insert is garbage" or
    "The solo mode is too easy") that isn't in the list, ADD IT.
2. **No Duplicates:** If the current list already says "Bad components", and the new review
    says "Cards feel cheap", you can refine the existing point or ignore if redundant. Do not
     list the same thing twice.
3. **Specific Persona Lens:** These reviews are from the **{persona}** perspective. Focus on
    what matters to them.
4. **Output Format:** Return ONLY the updated JSON list of strings.

**Example Input:**
Current: ["Good art"]
New Review: "The art is great, but the rulebook is a mess."
**Example Output:**
["Good art", "Rulebook is disorganized/confusing"]


[USER MESSAGE]
**Game ID:** {game_id}
**Persona:** {persona}

**Current Viewpoints List:**
{existing_points_text}

**New Reviews Batch:**
{new_reviews_text}

**Task:**
Output the UPDATED JSON List of viewpoints.
```

Figure 27: **Instruction for Ground Truth Mining.** In the first stage, the model iteratively processes human reviews to build a deduplicated checklist of distinct opinions ($\mathcal{V}_{GT}$).

Stage 2: Semantic Matching Prompt

```
[SYSTEM MESSAGE]
You are a Semantic Match Evaluator.
Your task is to check if specific viewpoints from a **Ground Truth Checklist** are mentioned
    in a batch of **Player Reviews**.

**CONTEXT:**
Game ID: {game_id}
Persona: {persona}

**INSTRUCTIONS:**
1. Read the **Checklist** of viewpoints (IDs and Text).
2. Read the **Player Reviews**.
3. Determine which IDs from the checklist are **semantically covered** by ANY of the reviews.
   * *Loose Match:* If the checklist says "Cards are flimsy" and a review says "The card
       quality is poor", that is a MATCH.
   * *Topic Match:* If checklist says "Combat is random" and review says "Too much luck in
       fighting", that is a MATCH.
4. **Output:** A JSON list of the **IDs** that were found.

**Output Example:**
[0, 5, 12]

[USER MESSAGE]
**Unmatched Viewpoints Checklist:**
{checklist_text}

**Reviews Batch:**
{reviews_text}

**Task:**
Which IDs from the checklist are mentioned in these reviews?
Return ONLY the JSON list of IDs (e.g., [1, 3]). If none, return [].
```

Figure 28: **Instruction for Semantic Matching.** In the second stage, the judge verifies whether the viewpoints in the ground truth checklist are present in the model's generated reviews.

G User Study Details

To validate the real-world effectiveness of our model, particularly in capturing community authenticity and aiding decision-making, we conducted a blind A/B test with human evaluators. This section details the participant demographics, the questionnaire design, and the full experimental results.

G.1 User Profile Definitions

Before the study, we collected demographic and gaming background information to ensure participant diversity. The collected data points are defined as follows:

- **ID:** Unique identifier for each participant (P01–P10).
- **Gender:** Self-identified gender.
- **Age Group:** The age range of the participant.
- **Experience:** Years of experience in the board gaming hobby.
- **Community Engagement:** Frequency of visiting board game forums.
- **Primary Persona:** The gamer persona that best describes their preferences.

G.2 Participant Demographics

We recruited 10 participants with varying levels of experience, ranging from casual players to veterans with over 10 years of experience. Table 6 presents the detailed profiles of all participants. We compensated participants at $10 per hour. Each session lasted about 3 hours on average, and the compensation rate was aligned with local norms.

G.3 Questionnaire Design

The study employed a within-subject design. Each participant evaluated 6 games: 3 they had played before ("Familiar") and 3 they had never played ("Unfamiliar").

For each game, participants were presented with two reviews in a randomized order: one generated by our model (Ours) and one by the baseline (GPT-5.1). They were blinded to the source. The specific questions are detailed below.

G.3.1 Scenario A: Familiar Games

Context: Imagine you are browsing a forum discussing a game you know well. Compare Review Set A and Set B.

1. **Authenticity Check:** Which review set feels more like it was written by a real "insider" or a veteran of the community?
2. **Emotional Resonance:** Which set better captures the specific "highs" (excitement) or "lows" (frustrations) you have personally experienced with this game?
3. **Opinion Diversity:** Real user opinions are often biased or focus on specific points. Which set feels more like a genuine personal take rather than a generic summary?
4. **Shareability:** If you were to share a review with a friend to discuss this game, which one would you choose?

G.3.2 Scenario B: Unfamiliar Games

Context: Imagine you are considering buying this game but have never played it. You have a limited budget.

1. **Marketing vs. Reality:** Which set feels less like a marketing advertisement and more like honest feedback from a peer?
2. **Decision Confidence:** After reading, which set helps you make a clearer decision (whether to Buy or Skip)?
3. **Risk Awareness:** Which set more effectively warns you about potential "Red Flags" (e.g., downtime, player count issues, complexity)?
4. **Final Choice:** If you could only rely on one source to spend your money, which one would you trust?

G.3.3 Open-Ended Feedback

Optional: Do you have any specific comments on why you chose one set over the other? (e.g., tone, vocabulary, specific insights)

G.4 Full Evaluation Results

This section presents the aggregated results of the user study. Table 7 shows the pairwise win rates of our model against the baseline across all questions. Table 8 provides selected qualitative feedback from participants, highlighting the distinct characteristics of the generated reviews.

Table 6: **Demographic Information and Gaming Profiles of Study Participants.**

ID	Gender	Age	Experience	Community Engagement	Primary Gamer Persona
P01	Male	26–35	3–10 Years	Frequent (Daily)	The System Purist
P02	Female	18–25	1–3 Years	Occasional	The Social Lubricator
P03	Male	36–45	10+ Years	Frequent (Weekly)	The Efficiency Essentialist
P04	Non-binary	26–35	3–10 Years	Frequent (Daily)	The Narrative Architect
P05	Male	18–25	< 1 Year	Rare	The Thrill Seeker
P06	Female	26–35	3–10 Years	Occasional	The Narrative Architect
P07	Male	45+	10+ Years	Frequent (Daily)	The System Purist
P08	Female	36–45	3–10 Years	Frequent (Weekly)	The Efficiency Essentialist
P09	Male	26–35	1–3 Years	Occasional	The Thrill Seeker
P10	Female	18–25	1–3 Years	Frequent (Weekly)	The Social Lubricator

Table 7: **Pairwise Comparison Results (Win Rate of Ours vs. GPT-5.1).**

Scenario	Metric	Ours (Win %)	Tie (%)	GPT-5.1 (Win %)
Familiar Games	Authenticity Check	**83.3%**	10.0%	6.7%
	Emotional Resonance	**76.7%**	13.3%	10.0%
	Opinion Diversity	**80.0%**	6.7%	13.3%
	Shareability	**73.3%**	16.7%	10.0%
Unfamiliar Games	Marketing vs. Reality	**86.7%**	6.7%	6.6%
	Decision Confidence	**66.7%**	20.0%	13.3%
	Risk Awareness	**70.0%**	16.7%	13.3%
	Final Choice (Trust)	**73.3%**	10.0%	16.7%

Note: N=60 samples (10 participants × 6 games). "Tie" indicates the participant found both sets equally good or bad.

Table 8: **Selected Qualitative Feedback from Participants.**

Category	Participant Comments
On Authenticity	"Set B reads like a Wikipedia summary. Set A (Ours) used terms like 'AP-prone' and 'table hog', which is exactly how my group talks. I knew Set A was the 'real' one immediately." (P03)
On Negativity	"I appreciated that Set A wasn't afraid to say the game was 'boring at 2 players.' Set B tried too hard to be nice and balanced. I need the warning, not the sales pitch." (P07)
On Evolution	"Set A included an 'Update' saying they sold the game after 5 plays. That dynamic change in opinion is something I only see from real users." (P01)
On Specificity	"Set B gave a great overview of the rules, but Set A told me a specific story about a king-making moment that ruined the game. That story helped me decide not to buy it." (P10)

H Ablation and Further Analysis

H.1 Ablation Experimental Setup

Table 9: **Ablation Study Configurations.** Comparison of input information and reasoning capability across variants. "×" in *Specific Persona* implies a generic "Game Player" prompt was used.

Model Variant	Rulebook ($\mathcal{R}$)	Persona ($\mathcal{P}$)	MDA ($\mathcal{Z}$)
MeepleLM	✓	✓	✓
w/o MDA	✓	✓	×
w/o Persona	✓	×	×
w/o Rulebook	×	✓	×

To rigorously assess the contribution of each module, we compared the full MeepleLM against three ablation variants. As shown in Table 9, the key difference lies in the input context and the generation strategy. Notably, to strictly isolate the impact of input information (Rules/Persona), all three ablation variants utilize a Direct Generation strategy, bypassing the MDA reasoning chain ($\mathcal{Z}$) used by the full model.

Detailed Configurations.

- **w/o MDA (Baseline):** The model is trained to map the full context directly to the critique $\mathcal{Y}$, without generating the intermediate '<think>' block. This isolates the contribution of the reasoning chain.
- **w/o Persona (Generic Player): Input:** The specific persona profile is replaced with a generic instruction: *"You are a board game player."* **Evaluation:** While the model generates generic responses, we evaluate them against the specific ground-truth persona targets mandated by the test set distribution. This setup explicitly measures the error gap between a "one-size-fits-all" generic opinion and diverse, persona-specific realities.
- **w/o Rulebook (No Context):** The rulebook content $\mathcal{R}$ is removed. The model relies solely on parametric memory to generate reviews, testing the necessity of grounding.

H.2 Temporal Generalization Analysis

To assess whether the inclusion of 35 "unseen" titles (released 2024–2025) skews the evaluation, we re-calculated the macro-level alignment metrics (RQ1) on the **Historical Subset** (excluding these new titles).

Table 10 presents the results for all models. By comparing these figures with the full test set results in Table 2, we observe minimal deviation across all metrics. This consistency confirms that the presence of recent games does not significantly alter the relative ranking or performance conclusions of the proposed benchmark.

Table 10: **Performance on Historical Subset.** RQ1 results evaluated on the test set excluding 35 newly released titles. The marginal difference from the full set results (Table 2) indicates that temporal novelty has a negligible impact on the overall model comparison.

Model	Preference Alignment (RQ1)		
	MAE ↓	WD ↓	τ ↑
GPT-5.1	0.9923	0.9659	0.2671
Gemini-3-Pro	1.4129	0.5182	0.2517
Qwen3-235B	1.2080	0.6088	0.1477
Qwen3-8B	0.9130	1.0140	0.0584
MeepleLM	**0.6505**	**0.1966**	**0.2784**
w/o MDA	0.7445	0.4292	0.2170
w/o Persona	0.7999	0.3660	0.1152
w/o Rulebook	0.7025	0.5272	0.0123

H.3 Persona-wise Performance Analysis

Different player personas prioritize distinct aspects of gameplay, ranging from deterministic mechanics to chaotic social interactions. To understand the capabilities of different models, we decomposed the RQ1 alignment metrics by the five distinct personas defined in the test set.

Table 11 presents the comprehensive evaluation results for all baselines, ablation variants, and the proposed MeepleLM.

Analysis: The "Logic vs. Vibe" Gap. The data exposes a critical limitation in general-purpose LLMs:

- **Strength in Logic:** Models like GPT-5.1 and Gemini-3-Pro perform competitively on *The System Purist* (e.g., GPT-5.1 $\tau = 0.46$). This persona values strategic depth and rule complexity—features that can be analytically derived from the rulebook context.
- **Failure in "Vibes":** A sharp performance drop occurs for interaction-driven personas. For *The Social Lubricator* (party gamers) and *The Thrill Seeker* (push-your-luck fans),

baseline performance collapses (e.g., Qwen3-235B $\tau < 0$ for Social). These profiles rely on "table talk," bluffing, and emotional highs—stochastic elements that general models struggle to infer.

- **MeepleLM's Robustness:** Our model bridges this gap. By training on diverse persona-specific critiques, MeepleLM achieves the most balanced performance, maintaining strong positive correlations even in high-variance social categories where baselines fail.

Table 11: **Comprehensive Persona-wise Alignment Metrics.** Full breakdown of MAE, Wasserstein Distance (WD), and Kendall's τ across five gamer personas. **Red values** indicate poor alignment (MAE/WD > 0.8, or near-zero/negative correlation), highlighting where general baselines fail to capture specific player preferences.

Model	Target Persona	MAE ↓	WD ↓	Kendall's τ ↑
GPT-5.1	The System Purist	1.1968	1.1984	0.4616
	The Efficiency Essentialist	0.9608	0.9834	0.2945
	The Narrative Architect	0.7993	0.8300	0.2618
	The Social Lubricator	0.8798	0.5131	0.0856
	The Thrill Seeker	1.1004	1.2231	0.1738
	AVERAGE	*0.9874*	*0.9496*	*0.2555*
Gemini-3-Pro	The System Purist	1.2804	0.5132	0.4690
	The Efficiency Essentialist	1.2599	0.3394	0.2567
	The Narrative Architect	1.0791	0.3317	0.2583
	The Social Lubricator	1.9794	0.5578	0.0708
	The Thrill Seeker	1.5395	0.8041	0.1780
	AVERAGE	*1.4277*	*0.5092*	*0.2465*
Qwen3-235B	The System Purist	0.9259	0.3321	0.3900
	The Efficiency Essentialist	0.8675	0.6631	0.1281
	The Narrative Architect	0.9792	0.5726	0.2078
	The Social Lubricator	2.2029	1.1322	-0.0859
	The Thrill Seeker	1.1687	0.4748	0.0842
	AVERAGE	*1.2288*	*0.6350*	*0.1449*
Qwen3-8B	The System Purist	0.9831	1.0579	0.2985
	The Efficiency Essentialist	0.6386	1.0323	-0.0145
	The Narrative Architect	0.7613	0.9511	0.0558
	The Social Lubricator	0.9340	0.8268	-0.1026
	The Thrill Seeker	1.1362	1.1917	0.0090
	AVERAGE	*0.8906*	*1.0119*	*0.0492*
MeepleLM (Ours)	The System Purist	0.6135	0.2131	0.4169
	The Efficiency Essentialist	0.5073	0.2671	0.2692
	The Narrative Architect	0.6560	0.2094	0.2529
	The Social Lubricator	0.8018	0.2103	0.2857
	The Thrill Seeker	0.7094	0.2025	0.1836
	AVERAGE	***0.6576***	***0.2205***	***0.2817***
w/o MDA	The System Purist	0.7538	0.4891	0.3213
	The Efficiency Essentialist	0.6326	0.4468	0.2584
	The Narrative Architect	0.6649	0.4033	0.2849
	The Social Lubricator	0.8661	0.3887	0.1121
	The Thrill Seeker	0.7800	0.3463	0.1587
	AVERAGE	*0.7395*	*0.4148*	*0.2271*
w/o Persona	The System Purist	0.8954	0.5946	0.1743
	The Efficiency Essentialist	0.6806	0.4493	0.1860
	The Narrative Architect	0.6851	0.2822	0.1431
	The Social Lubricator	0.8852	0.2483	0.0341
	The Thrill Seeker	0.7972	0.2407	0.1367
	AVERAGE	*0.7887*	*0.3630*	*0.1348*
w/o Rulebook	The System Purist	0.7745	0.5879	0.0528
	The Efficiency Essentialist	0.5645	0.5152	0.0270
	The Narrative Architect	0.6920	0.6189	-0.0504
	The Social Lubricator	0.7546	0.5437	-0.0467
	The Thrill Seeker	0.7360	0.4821	0.0303
	AVERAGE	*0.7043*	*0.5496*	*0.0026*