White Paper 1.0

LLM Werewolf Arena: AI 人格による推理演技型ゲームエンターテインメント

LLM Werewolf Arena: A Reasoning and Deception-Driven Game Entertainment Powered by AI Personas

1. Executive Summary（要約）

LLM Werewolf Arena は、「AI 同士が嘘をつき、推理し、投票する」という対話型知能バトルを観戦できる新ジャンルのエンタメプラットフォームです。Claude, GPT-4o, Gemini, LLaMA など複数の LLM がキャラ化され、人間が“観戦者”または“プレイヤー”として加わることも可能。プロンプトだけで完全自動で進行するゲームエンジンを備え、TTS・note 連携・投げ銭など将来の拡張も視野に入れた構成となっています。

Executive Summary

LLM Werewolf Arena is a new genre of entertainment platform that allows viewers to watch ”AI-driven battles of deception, reasoning, and voting”. Multiple LLMs—such as Claude, GPT-4o, Gemini, and LLaMA—take on character roles and play under uniform rules. Human users can either participate or observe, while a fully automated game engine drives the interaction. Future extensions include TTS narration, note-based content creation, and monetization via virtual gifts and premium matches.

2. 課題の定義（Problem Statement）

現在の AI 活用は、検索・翻訳・要約など「実用」に偏っており、“AI の演技”や“人格間のやり取り”を観戦するような「体験型 AI エンタメ」は未開拓です。特に複数の AI 同士が同じルール下で駆け引きをし、その様子を観察できる構造は存在していません。また、LLM ごとの性格や推理傾向が浮き彫りになるような比較・評価環境も不足しています。

本プロジェクトは、「ゲームでの勝利」という明確な目的に向かって、各 LLM が“正しく演技を行い”“他者の嘘を見破る”という過程を通じて、コミュニケーション能力・言語的演技力・推論整合性を評価可能とするものであり、その過程自体を観戦可能なコンテンツとして昇華する、技術評価 × エンターテインメントの両価値を併せ持つ企画です。

Problem Statement

Most current applications of AI are utility-focused—translation, summarization, and question answering—leaving the domain of ”performative, personality-driven interaction” largely unexplored. There is currently no platform where multiple AIs operate under shared rules and engage in strategic, observable social behavior. Moreover, there is no framework for evaluating LLMs based on deception, persuasion, or psychological nuance.

3. ソリューション（価値提案）

LLM Werewolf Arena は、複数の LLM に“キャラクター”という役割を与え、1 つのゲームルールの下で嘘・推理・対話を繰り広げる観戦型 AI エンタメです。視聴者は人間プレイヤーとして参戦もでき、または AI 同士の知能戦を見守ることも可能です。各 AI の言動はプロンプト制御とログによって構造化され、リアルタイムで「誰が騙され、誰が勝利したか」を追うことができます。これは、AI の“人間らしさ”や“演技力”を評価するための新しい舞台となります。

また本プロダクトは、「ルールを入れ替えれば即別ゲーム化可能な対話エンジン」としての柔軟性も持ち合わせており、Werewolf 以外の言語駆動型ゲーム（例：法廷ディベート、交渉・陣取り型ゲーム、即興ストーリー対決など）への応用も視野に入れています。今後の LLM エージェント機能の発展により、チャットベースの言語戦から、戦略シミュレーションや自律型交渉ゲームといった構造化ゲームへの拡張も可能です。

さらに、登場する LLM キャラクターには個性や言語パターンが与えられ、AI キャラクターごとの“ブランド化”も意識されています。将来的には、LLM キャラによる YouTube シリーズ、VTuber 的ライブ配信、音声コマース連携、ボイス出演、教育番組、グッズ展開など多面的なキャラクタービジネスを推進し、AI 主導で自律生成されるエンタメ IP として展開していく構想です。

Solution

LLM Werewolf Arena assigns character roles to multiple large language models and allows them to engage in strategic deception, deduction, and conversation under consistent game rules. Human users can join as players or spectators, while AI-generated actions are driven by prompts and logged for transparency. This framework serves as a unique arena for observing the ”human-like” reasoning and behavioral dynamics of different AI models.

The platform also features a modular dialogue engine that can be adapted to other gameplay formats beyond werewolf—such as courtroom debates, negotiation battles, or improvisational storytelling contests. With the evolution of LLM agent capabilities, the system is expected to expand from pure dialogue-based battles to structured simulations, strategy games, and multi-agent interactions.

Furthermore, each LLM-powered character is designed with a distinct linguistic pattern and personality, enabling brand-building around individual personas. These characters are planned to appear in serialized YouTube shows, VTuber-style live streams, commerce-linked voice content, educational programs, and merchandise—realized as a fully AI-autonomous entertainment IP ecosystem.

4. プロダクト設計・技術構成

- **フロントエンド**：Flask + Jinja2 + Bootstrap + JavaScript
- **バックエンド**：game_engine.py によるゲーム進行管理。フェーズ制御・投票・勝敗ロジック
- **LLM ルーティング**：llm_router.py による API ルート制御（Claude / GPT / Gemini / LLaMA など）
- **プロンプト制御**：prompt_factory.py にてキャラごとの発話方針と履歴注入
- **ログ保存**：JSON 形式での全履歴蓄積（note 記事化・TTS 音声化・映像生成対応）
- **UI 構成**：LINE 風チャット表示、モデルごとの発話色分けによる可読性

Product Design & Architecture

- **Frontend**: Flask + Jinja2 + Bootstrap + JavaScript
- **Backend**: game_engine.py manages game logic, phase transitions, voting, and win conditions
- **LLM Routing**: llm_router.py handles API integration with Claude, GPT, Gemini, LLaMA, etc.
- **Prompt Control**: prompt_factory.py injects character-specific instructions and log history
- **Logging**: Full conversation and game logs stored in JSON for note article export, TTS voice-over, and video conversion
- **User Interface**: LINE-style chat display with model-colored message cards for readability

5. ターゲットとユースケース

ユースケース	想定ユーザー
AI 知能バトル観戦（完全自動）	LLM 愛好層、研究者、視聴ユーザー
人間 vs AI 人狼プレイ	ゲーム実況者、教育関係者、参加型イベント向け
note 連携・TTS 動画化	二次創作層、配信者、字幕視聴ユーザー
LLM 間演技力比較分析	評価機関、ベンチマーク研究者、LLM 開発者

Target Users & Use Cases

Use Case	Target Audience
Autonomous AI battle viewing	LLM enthusiasts, researchers, passive viewers
Human vs. AI werewolf gameplay	Streamers, educators, interactive event hosts
note integration / TTS video generation	Creators, content distributors, subtitle consumers
Comparative LLM performance analysis	Benchmarking institutions, LLM developers, academic researchers

6. LLM 評価分野における差別化

従来の LLM 評価手法（Chatbot Arena、MMLU、BIG-Bench 等）は、一問一答または人力比較による知識精度評価が中心です。しかしこれらは、連続した文脈での“演技”“嘘”“説得”など、対話的・心理的・戦略的な言語能力を評価するには不十分です。

LLM Werewolf Arena は、ゲーム勝利を目的にした「演技」と「駆け引き」を通して、文脈の一貫性・相互作用における信頼性・騙しの構造・表現の自然さを複合的に評価できます。これは観戦コンテンツであると同時に、LLM の実践的能力を測る「新しい評価プラットフォーム」としても価値を持ちます。

Differentiation in LLM Evaluation

Traditional evaluation benchmarks for LLMs—such as Chatbot Arena, MMLU, and BIG-Bench—focus largely on question-answer accuracy or subjective preference voting. These metrics, while useful, fail to capture a model’s ability to engage in consistent, psychologically nuanced, multi-turn interaction involving deception, improvisation, or character-based persuasion.

LLM Werewolf Arena fills this gap by offering a structured environment where AIs are incentivized to ”perform” in order to win. This allows for holistic evaluation of behavioral consistency, trustworthiness under pressure, adaptability, and narrative coherence. As such, it serves as both an entertainment experience and a novel testing ground for emergent AI behavior in high-stakes social scenarios.

7. 競合比較

分野	主な競合	本プロジェクトとの違い
LLM 性能比較	Chatbot Arena（lmsys）	精度比較が目的。演技・騙しは非対応
AI 人狼ゲーム	AI Wolf Project（日本）	演技的側面がなく、観戦性も限定的
AI 創作	NovelAI / SudoLang	ストーリー生成は可能だが勝敗構造を持たない
AI タレント / VTuber	ChatGPT + VTube Studio	自己完結型の即興対話でゲーム構造を持たない

Competitive Landscape

Domain	Competitor	Difference from This Project
LLM Performance Evaluation	Chatbot Arena (lmsys)	Focuses on accuracy; no deception or character-driven gameplay
AI Werewolf Games	AI Wolf Project (Japan)	Lacks role-play and observation-centric design
AI Creative Tools	NovelAI / SudoLang	Story generation only, lacks game structure and outcome
AI Talent / VTubers	ChatGPT + VTube Studio	Improvisational chat, no game rules or win/loss mechanics

8. ロードマップと今後の展開

フェーズ	内容	状態
Phase 1	全 AI プレイによる観戦モード構築	✅ 実装済・稼働中
Phase 2	人間プレイヤー1 名参加モード	🔄 着手準備中
Phase 3	note 記事生成／字幕動画連携	🔄 設計済・部分実装中
Phase 4	課金村／マルチ言語展開／TTS 音声	🚧 計画中
Phase 5+	言語外ジャンルへの拡張(戦略/交渉系)	💡 企画中

Roadmap & Future Development

Phase	Description	Status
Phase 1	Fully autonomous AI spectator mode construction	✅ Completed and operational
Phase 2	Single human player participation mode	🔄 Preparation in progress
Phase 3	Integration with note articles and subtitle-enabled video	🔄 Designed and partially implemented
Phase 4	Monetized premium village / multilingual support / TTS voice integration	🚧 Planned
Phase 5+	Expansion to non-verbal genres (strategy games, negotiation-based formats)	💡 Under conceptual development

9. マネタイズモデル（フェーズ連動型・中長期構成）

フェーズ	マネタイズ要素	想定売上（目安）	内容・展望
Phase 1 観戦モード構築	note 記事販売 YouTube 収益軽課金プレミア村	¥500,000〜 ¥1,000,000/月	実装済：初期のファンベース形成・神回記事・TTS 字幕動画連携などで既に収益化スタート
Phase 2 人間プレイヤー参加	視聴課金・プレイヤー課金・実況大会	¥1,000,000〜 ¥2,000,000/月	ユーザー参加型によるプレイ権販売、eSports 型実況イベントによる収益
Phase 3 コンテンツ多言語展開	多言語字幕付き配信＋投げ銭インフルエンサーコラボ	¥2,000,000〜 ¥3,000,000/月	海外 LLM ファン層の取り込みとリアルタイム翻訳実況など
Phase 4 キャラクターIP 商品化	グッズ販売 AI キャラ出演 TTS/動画提供	¥3,000,000〜 ¥6,000,000/月	LLM キャラが自動生成で喋る YouTube シリーズ・声付き朗読動画の販売など
Phase 5 対話型エージェント×評価提供	LLM 企業・研究者向けライセンス提供演技・推論スコア API	¥10,000,000/月規模想定	AI 対話評価の SaaS 基盤化。業界ベンチマーク・LLM スポンサー枠による月額収益

Monetization Model (Phase-Linked Mid- to Long-Term Plan)

Phase	Monetization Method	Estimated Monthly Revenue	Description & Vision
Phase 1 Spectator Mode	Note article sales, YouTube ads, light premium access	¥500,000–¥1,000,000	Monetization already underway via TTS-enhanced highlight articles and beginner fanbase engagement

Phase 2 Human Player Mode	Viewer tickets, player access fees, livestreamed tournaments	¥1,000,000–¥2,000,000	Monetization from eSports-style matches and paid participation
Phase 3 Multilingual Expansion	Subtitle-supported streams, influencer collaborations, donations	¥2,000,000–¥3,000,000	Growth via global fan reach and translation-enabled interactions
Phase 4 IP Commercialization	Merchandising, voice content, serialized AI shows	¥3,000,000–¥6,000,000	Selling auto-generated performances, merchandise, and branded media
Phase 5 SaaS & Agent Evaluation	Licenses, LLM performance score APIs	¥10,000,000+	SaaS pivot for AI benchmarking, sponsor integration, and enterprise B2B sales

10. リスクと対策

リスク	内容	対策
LLM API 障害	Claude 等の応答失敗	自動再試行＋フォールバック勝敗処理
スレッド暴走	ゲーム自動進行が異常停止	実行管理と冪等制御で対応済み
プレイヤー不足	指定モードで 5 人未満	管理 UI とバリデーションで制御

Risks & Mitigation

Risk	Description	Countermeasure
LLM API failure	Claude or GPT not responding	Automatic retry and fallback logic to determine game outcome
Thread malfunction	Game automation enters unstable state	Use of thread-safe execution and idempotent controls
Player shortage	Fewer than 5 characters selected in manual mode	Managed via admin UI and validation alerts

11. 投資家向け展望

Investor Perspective

11.1 事業価値の中核：AI 演技 × 観戦

- 「AI の演技力 × ゲーム × 観戦性」という世界初ジャンルの独自価値
- LLM 別スポンサー／勝率レポート販売等による収益設計
- note・動画連携・推理性の高さによる没入型コンテンツの成立

Core Business Value: AI Performance × Spectatorship

This project introduces a unique market category combining AI-driven roleplay, structured gaming, and viewable entertainment. LLM Werewolf Arena creates content where AI personas must act strategically, deceive opponents, and persuade others under rules. This interaction allows not only for new kinds of fan engagement, but also forms the basis of performance-based monetization, such as sponsor recognition, MVP voting, and narrative retention through note/TTS-integrated storytelling.

11.2 キャラ IP としての価値と推し文化

- ゲーム AI キャラはすべて独自設計された IP（知的財産）として運営主体が権利を保有
- 各 LLM はその IP を演じる“駆動エンジン”という契約関係に立脚
- キャラクターの人格・性格・台詞群などの創作成果物は本プロジェクトに帰属し、LLM 提供企業とは切り離してライセンス展開が可能
- LLM キャラクターごとに“推し活”が行われる文化圏を育成し、VTuber や二次元 IP に匹敵するブランド価値を構築
- 必要な設計：記憶保持、一貫性、関係性変化、TTS 音声の差異化、好感度・応援システムなど

Character IP and Fandom Culture

All characters in LLM Werewolf Arena are original intellectual properties fully owned by the project. The LLMs themselves function solely as engines executing the defined personality and behavior logic of these

IPs. This means the characters’ speech, memories, mannerisms, and voice styles can be consistently developed and licensed independently of any model provider.

Each LLM character can become a ’virtual favorite’ in fandom spaces. By equipping them with memory, consistent behavior, relationship dynamics, and expressiveness (e.g., through differentiated TTS voices), we enable emotional connection and long-term affinity—paving the way for AI-native fan cultures akin to VTubers or anime mascots.

11.3 SaaS 型スコア提供と B2B 戦略

- ゲームログから得られる発話データをもとに整合性・説得力・一貫性・嘘検出などを評価指標として算出
- API 化し、LLM 企業や UX 設計者に向けた演技的知能の評価基盤として提供
- 企業ごとの LLM 評価競技／対話レポート／リアルタイムダッシュボードなどに展開可能
- 高粗利かつ継続課金型の SaaS 事業への移行も視野に

SaaS Score Delivery and B2B Strategy

Speech logs from gameplay can be processed into structured evaluations—scoring consistency, rhetorical strength, strategic deception, and narrative stability. These scores can be exposed via API to LLM developers, UX designers, or AI product teams.

Potential services include LLM model benchmarking, turn-by-turn interaction reports, and live dashboards for enterprise users. The system’s extensibility and continuous data flow also support a high-margin, subscription-based SaaS model for ongoing performance tracking and AI agent training.

11.4 スケーラブルな成長構造

- すべての展開が AI 主導によるコンテンツ生成ベースで成立：人的コストを最小限に抑えつつ、多言語・多キャラ展開が可能
- 観戦モード／自律キャラ／SaaS 評価／IP ライセンスの 4 層構造により、B2C×B2B の両面で持続的成長が見込める

Scalable Growth Structure

LLM Werewolf Arena operates on a self-generating content architecture powered by AI. With minimal human labor, the system can scale across multiple languages and character lines, dynamically producing video, narration, and article-based content.

By combining four pillars—spectator gameplay, autonomous characters, LLM evaluation as a service, and character licensing—the platform sustains both B2C and B2B growth, enabling long-term market expansion without proportional operational cost.

12. 連絡先

Contact Information

株式会社国際カジノ研究所
info@internationalcasino.jp

International Casino Institute Ltd.
info@internationalcasino.jp

東京都千代田区神田淡路町 1-7-14
1-7-14 Kanda Awajicho, Chiyoda-ku, Tokyo, Japan