AI Debate Arena: Designing a Persona-Driven LLM Benchmark (2025)

Introduction

Building a **serious LLM benchmark around structured, persona-driven debates** requires integrating proven practices from modern evaluations. As of **late 2025**, leading benchmarks assess models on multifaceted criteria – from reasoning and factual accuracy to persuasion and logical coherence – using both human and AI judgments. This report distills current benchmarking methodologies and proposes an updated **AI Debate Arena** that is both engaging (crowd-powered debates) and scientifically rigorous. We will examine how today's benchmarks operate, how leaderboards handle model entries and scores, best practices for crowd interaction, which frontier models to include, and how to implement a dual scoring system combining **user votes and LLM-judged metrics**. Finally, we outline a concrete plan to elevate AI Debate Arena into a credible **entertainment-plus-research** platform.

1. Evaluation Metrics and Methods in Modern LLM Benchmarks (2025)

Contemporary LLM benchmarks use a blend of **qualitative judgments and quantitative tasks** to evaluate models. Key evaluation dimensions include:

- **Reasoning & Logic:** Ability to perform multi-step reasoning, maintain coherence, and avoid fallacies. Benchmarks like BigBench Hard (BBH) and newer "adversarial" reasoning tests (e.g. *MuSR* for multi-step puzzles) directly measure logical problem-solving [1] [2]. Judges (human or AI) also rate conversational logic – e.g. whether an argument is internally consistent and well-structured [3] [4].

- **Factual Accuracy:** How truthfully and accurately a model responds. Standard QA benchmarks (MMLU, specialized QA sets like *GPQA*) gauge knowledge across domains [5] [6]. In debates, factual accuracy can be evaluated by checking each side's claims against references or via an LLM judge that flags incorrect statements. Ensuring factuality is crucial so that a persuasive but incorrect argument doesn't "win" unfairly.

- **Persuasion & Rhetorical Skill:** The effectiveness in convincing or engaging a reader. This is harder to quantify with static tests, so **human preference votes or LLM-as-a-judge** evaluations are used. For example, the **Debate Speech Evaluation** benchmark requires judges to assess persuasiveness, argument strength, and style in long-form debates [3] [4]. Metrics here are often qualitative (which speech is more convincing or better framed) and may rely on **LLM judges** trained or prompted to score rhetoric.

- **Knowledge & Coverage:** Breadth of world knowledge and ability to answer diverse questions. Benchmarks like MMLU-Pro (an updated multitask exam) cover professional-level knowledge with

tougher questions [6]. High scores on such exams (e.g. >80% on MMLU-Pro or comparable professional quizzes) indicate strong factual coverage. Debates can indirectly test this by picking topics from various fields and seeing if models bring in relevant factual content.

- **Following Instructions & Formatting:** Some evaluations specifically test if models follow instructions to the letter (important in structured tasks). For instance, *IFEval* measures a model's adherence to explicit instructions or formats [7]. In a debate setting, this could translate to sticking to debate rules or personas (not breaking character).

Scoring Approaches: Modern benchmarks employ both **human and AI judges**, and scoring can be done **per round or for the full exchange** depending on the format: - **Human Judgment (Crowd or Experts):** The **LMSYS Chatbot Arena** exemplifies crowd-based evaluation: two models chat and a user votes on the better response, with aggregated votes producing an Elo-style rating [8] [9]. Humans naturally consider overall persuasiveness and quality of the entire dialogue when voting (effectively judging the full debate rather than isolated turns). In some cases, expert panels are used for specific metrics (e.g. domain experts rating correctness of answers in a medical AI benchmark). Human scoring is often considered the gold standard for subjective criteria like helpfulness or style.

- **AI as Judges:** Increasingly, strong LLMs (like GPT-4.0/4.1 or later) are used to **automatically evaluate model outputs**, either to supplement or approximate human ratings. Research in 2023–2025 found that GPT-4-based judges agree with expert evaluations ~80% of the time on dialogue quality [10]. Frameworks such as **MT-Bench** use GPT-4 to score multi-turn conversations on criteria like relevance, reasoning, and correctness, yielding a numeric "MT-Bench score" for each model [11]. Similarly, *AlpacaEval* and others have pioneered using an LLM to decide which of two answers is better. These AI judges can provide **round-by-round feedback** (e.g. scoring each response turn) or a **final verdict** for a whole dialogue. In practice, most benchmarks use them for an overall judgement per comparison (simulating a human vote). For debates, one could employ an LLM judge to evaluate each round's winner *and* the debate's overall winner, though consistency of the judging model becomes a consideration (to avoid, for example, a bias toward whoever spoke last – a known *positional bias* in some AI judges [12]).

- **Round-by-Round vs. Full-Dialogue Scoring:** Traditional QA benchmarks score individual questions (akin to round-by-round), but in open-ended debates it's more common to score the **overall outcome**. Chatbot Arena, for instance, lets a user read the entire multi-turn debate and then cast a single vote for the winner [13]. This holistic scoring captures which model ultimately made the more compelling case. However, a more granular approach can be useful: one could assign points each round (e.g. "Model A clearly rebutted Model B in round 2") and tally them. This *round-by-round scoring* could surface if a model consistently starts strong but falters in later rebuttals, etc. Currently, no major public benchmark scores every turn explicitly – they focus on the final preference – but designing the Arena, we might combine both: e.g. have an LLM judge award *round points* while the human/crowd vote decides the final winner. This dual view could enrich the analysis of debate dynamics.

- **Evaluation Frameworks:** A variety of frameworks exist to facilitate these evaluations. **Stanford's HELM** (Holistic Evaluation of Language Models) provides a structured approach to test models across scenarios and metrics, and encourages *prompt-level transparency* (detailed per-question results) [14]. The **Eleuther AI LM Evaluation Harness** underpins the Hugging Face Leaderboard, enabling

consistent testing on many tasks [15]. **OpenAI Evals** (an open framework by OpenAI) allows user-created evaluation scripts for custom metrics. For dialogue and debate specifically, academic proposals like **Debate Speech Evaluation** (which we saw above) and tools like **JudgeBench** focus on rigorously testing LLMs' judging capabilities [12]. We can leverage these frameworks – for example, use LM Evaluation Harness for standardized quiz tasks, and a ChatArena-style setup for debates – to ensure our benchmark covers both **objective correctness and subjective quality**.

In summary, state-of-the-art benchmarks employ a **mix of metrics**: standardized test scores for knowledge and reasoning, plus **human or AI-driven preference scores** for qualities like helpfulness, coherence, and persuasion. This mixed methodology will inform our debate arena scoring (combining objective and subjective assessments).

2. Leaderboard Structures in High-Traffic Benchmarks

High-profile leaderboards in 2025 (LMSYS's **Chatbot Arena**, Hugging Face's **Open LLM Leaderboard**, Stanford's **HELM**, etc.) have converged on certain best practices in how they rank and present models:

- **Anonymous Battle Ranking (LMSYS Chatbot Arena):** The LMSYS **Chatbot Arena** popularized a *crowdsourced Elo leaderboard*. Users are presented with side-by-side answers from two unnamed models and vote for the better output [13]. Each win/loss updates the models' Elo scores (derived via a Bradley–Terry statistical model) with **confidence intervals** shown to indicate uncertainty [8] [16]. This results in a live leaderboard of models ranked by Elo-like score, often with **error bars** to reflect rating confidence [17] [18]. Arena's leaderboard is segmented by category: originally text-based chat, but by 2025 it expanded to multiple "arenas" – e.g. **Text, Vision, Coding, Multimodal** – each with its own rankings [19]. This segmentation acknowledges that a model's rank may differ by domain (a top text model might not be the top in coding, etc.). **Model versions** are treated as separate entries; for example, GPT-4, GPT-4 (2024 update), and GPT-5.1 would each have their own entry if available, so progress can be tracked historically. LMSYS allows a wide range of **open-source and closed models** to compete (they've included everything from open LLaMA variants to closed API models like Claude or GPT, via partnerships or user-provided access). Notably, many models are first added under **codenames** or anonymous IDs for testing – especially unreleased versions – and only get a public name on the board once they are officially released and have enough battles to get a stable rating [20] [21]. This approach ensures new entrants (or fine-tuned variants) can be vetted without immediately influencing the public rankings until they're proven. **Score explainability** on Arena is relatively simple: the Elo score is a single summary metric of win-rate adjusted for opponent strength. However, Arena augmented this by publishing the raw conversation dataset and describing their Elo methodology in a paper [22] [23]. Users can thus dig into sample battles or refer to the methodology for interpretation. The **Arena UI** also emphasizes transparency by revealing model identities *after* each vote (so voters learn which model they preferred, but only post-vote to avoid bias) [13].

- **Task-Based Aggregate Leaderboard (Hugging Face Open LLM):** Hugging Face's leaderboard takes a **benchmark suite approach**: each model is evaluated on a fixed set of tasks in a controlled environment, producing a composite score. As of 2025, the Open LLM Leaderboard uses **six key benchmarks** covering instruction following (IFEval), complex reasoning (BBH), math (MATH lvl5), advanced Q&A (GPQA), long-form reasoning (MuSR), and a difficult knowledge test (MMLU-Pro) [7] [6]. These results are combined (after normalization) into an overall score for ranking [24]. The

leaderboard is presented as a table where each model's score on each task can be expanded, offering more *explainability* than a single number. For example, one can see Model X's exact accuracy on MATH vs. its score on BBH, identifying strengths and weaknesses. Hugging Face only lists **open-source or open-access models** (requiring model weights or API to be available for automated testing), so proprietary models like GPT-4 appear only if someone releases an equivalent open version or a scored result. Each entry is tied to a specific model checkpoint/version; the platform allows filtering by model families or sizes. **Model versions** are handled by naming conventions and timestamps – e.g. `LLaMA-2 70B v1.0` vs `v1.1` could appear as separate entries if both are submitted. The leaderboard maintainers also provide an **FAQ and version history** explaining any metric changes or dataset updates [25] [26]. This ensures that if the scoring formula or tasks change (which happened with introduction of new benchmarks like GPQA), older scores are archived for reference. **Score explainability** is further aided by the open-source evaluation harness: users can reproduce the exact scores locally [27] [28], and detailed per-question logs are available for scrutiny in a public dataset [29]. In short, the HF leaderboard focuses on **reproducible, standardized metrics** and transparency about how the scores are computed – a complement to the more free-form Arena style.

- **Holistic Reports and Multi-Metric Comparison (Stanford HELM):** The **HELM** project at Stanford takes a **dashboard** approach rather than a simple rank list. HELM evaluates each model on a wide array of scenarios (e.g. summarization, dialog, extraction, etc.) and reports metrics for **accuracy, robustness, calibration, fairness, bias, and efficiency** among others. Instead of a single number or rank, HELM produces a report card for each model, and comparative charts for groups of models [14] [30]. Leaderboard "structure" here is more of a filterable interface where you can select models and metrics to compare. For example, one can see that Model A has higher knowledge accuracy but worse calibration than Model B. **Model versions** are distinctly labeled (with model, release date, and any fine-tuning info) – HELM deliberately includes older models (GPT-2, GPT-3, original BERT, etc.) up through the latest (GPT-4, Claude, etc.) to show **progress over time and baselines**. This historical inclusion is academically useful: you can track how far new models have surpassed the old. In 2025, HELM added an **Item Response Theory (IRT)** based methodology to better estimate model performance on adaptive difficulty questions [31] [32]. This means models face questions adjusted to their level, providing more statistically robust comparisons with confidence intervals (similar in spirit to Elo's confidence in Arena, but for static tasks). **Score explainability** is paramount: every metric is defined in the report, and you can drill down to per-item results (e.g. see exactly which questions a model got wrong). The trade-off is that HELM doesn't produce a flashy single "leader" – it emphasizes **multidimensional evaluation**. For our purposes, adopting some HELM principles means designing the debate benchmark to capture multiple axes (win rate, factual accuracy, etc.) and possibly publishing *detailed evaluations*, not just a winner's trophy.

Leaderboard Examples & Insights: To illustrate how these approaches manifest: - Chatbot Arena displays an Elo ranking with uncertainty bands and separates leaderboards for text, vision, and coding arenas [19] [33]. This multi-arena setup might inspire us to have categories in our debate arena (e.g. separate leaderboards for *philosophical debates*, *factual debates*, maybe *coding debates* if relevant). - The HF leaderboard provides credibility by listing benchmarks and linking to the code; any claim like "Model X is #1" can be traced to specific test results [34] [29]. We should similarly ensure any score in our arena can be backed by data (e.g. transcripts of debates, or at least aggregated stats). - Both LMSYS and HF stress **open access** and community contributions. LMSYS open-sourced their battle data and platform (FastChat) [22] [35], and HF's is built on open tools. For AI Debate Arena to be credible, adopting an **open data policy** (with

appropriate privacy safeguards) will be important. This might mean releasing a dataset of debate transcripts and outcomes periodically, or at least sharing summary statistics, so external researchers can validate findings.

In summary, high-traffic leaderboards tend to be **transparent, frequently updated, and careful about model entries**: - They uniquely identify model versions and often keep older models for context. - They include both open and closed models (Arena does both; HF focuses on open, but that's by design). - They provide *explainable scores*, either via confidence intervals (Arena) or metric breakdowns (HF, HELM). We will blend these ideas into the structure of AI Debate Arena's leaderboard.

3. Best Practices for User Interaction and Crowd Voting

Engaging the public as judges in a benchmark demands careful design to ensure the results remain valid and not gamed. Benchmarks that mix **entertainment with research** (like crowd-driven arenas) use several best practices:

- **Anonymize Model Identities:** To prevent brand bias or preconceived notions influencing votes, it's crucial that users do *blind* evaluations. LMArena (Chatbot Arena's 2025 incarnation) keeps model names hidden during the voting session – users only see "Model A" and "Model B" outputs – and reveals the identities *after* the vote is cast [13]. This way, a voter's preference is based on content quality, not on loyalty to GPT vs Claude, etc. This anonymity also deters **brigading** (where a community might rally to upvote a favored model regardless of output quality).

- **Randomized Pairings and Content:** The platform should pair models and serve debate topics in a way that neither model consistently gets an easy or hard topic. Random matchmaking and user-provided prompts ensure a variety of content. LMSYS even randomizes some stylistic aspects to avoid models being recognized; for instance, anecdotally they experimented with **style normalization** (making responses follow certain formats) to reduce tell-tale signs that could reveal the model and bias the voter [36]. Consistent random shuffling of which side is Pro vs Con (in a debate) or who speaks first can also mitigate any side advantages.

- **Encourage Diverse Participation:** A benchmark that draws only AI experts or only casual users could skew results. Successful crowd benchmarks cast a wide net – LMArena notes their community ranges "from AI enthusiasts and students to researchers, developers, and everyday users" [37]. By **keeping the interface simple and game-like**, it attracts users who might just be curious to play with AI, thereby increasing diversity. The *voting flow* should be quick and fun – e.g. one-click voting, immediate reveal of winners, maybe even gamified elements like personal stats for voters. Entertainment can be enhanced with features like sharing a particularly funny or interesting debate (some platforms allow users to share conversation transcripts post-vote). However, the design must ensure that while users are entertained, the data collected is useful (so maybe discourage frivolous single-word prompts or extremely short sessions – possibly by guiding users to provide meaningful debate topics).

- **Preventing Vote Gaming and Spam: Crowd-voting is susceptible to manipulation**, so robust back-end monitoring is needed. Known mechanisms include:

- *Rate limiting:* Limit how many votes a single user or IP can cast in a short time, to prevent one person from skewing results through sheer volume. (Arena's FAQ mentions that votes after identities are revealed do not count, and one user can vote many times but each prompt is fresh [38] [39]; presumably they also monitor for suspicious rapid voting.)
- *Anomalous pattern detection:* LMSYS has policies for detecting anomalous users [40] – e.g. if a user always votes for whichever side is a particular model beyond statistical expectation, or if a cluster of accounts all vote in sync, those votes might be down-weighted or filtered. While details aren't fully public (to avoid aiding cheaters), the presence of "anomalous user detection" is noted as an integrity measure [40].
- *Account or identity gating:* Some platforms require login or have a reputation system for voters. Even if open to all, having OAuth (Google/GitHub login) can deter bots. The privacy policy should clarify how user data (prompts, votes) is stored – LMArena's policy allows using votes for research but shares them anonymously [41] [42].
- *Goodhart's Law awareness:* When a leaderboard metric becomes high-profile, participants (or model creators) may try to "game" it. Tech press in 2024-25 reported concerns that labs were fine-tuning models specifically to excel in Arena battles ("benchmark tailoring") [43]. While we can't fully stop labs from optimizing for our benchmark (indeed, if they do, it indicates the benchmark's influence), we should *design the evaluation to be robust*: using a wide range of topics and personas so that overfitting to a narrow style won't guarantee wins. It's also wise to publicly address this – e.g. noting that the Arena is one signal and encouraging users not to chase the metric blindly [44].

- *Transparency and Data Access:* Providing researchers access to anonymized vote data (as LMArena does via datasets on HuggingFace [45]) allows external auditing. If someone suspects vote manipulation, they could analyze the released data for irregularities. This transparency can discourage would-be gamers, knowing that "many eyes" could catch anomalies.

- **Combining Crowd Votes with AI Judgments:** One particularly novel way to ensure crowd voting is *useful* (and not merely a popularity contest) is to **use AI judges in parallel** as a check. For example, if a large discrepancy arises between human votes and an LLM-judge's verdict (say humans favor Model A but the AI judge strongly favors Model B for factual accuracy), the organizers can review those cases. It might signal that users were swayed by style over substance, or possibly that the AI judge is picking up something humans don't. While we don't want to override human votes, surfacing such cases could be educational to users ("BTW, the AI referee noted several factual errors by the winning model that voters overlooked"). This approach turns potential "gaming" (e.g. a model using flashy but wrong answers to win votes) into a learning opportunity, and over time could even train the crowd to be more critical.

- **Community Guidelines and Moderation:** Blending entertainment means some users might input inappropriate or trivial prompts. Clear guidelines (no hate, no obviously fact-free "silly" debates if it skews data) and an option to remove certain battles from the official scoring (e.g. if someone tries a nonsense prompt that breaks both models) might be needed. A moderation system – perhaps AI-assisted – can filter out such cases or at least flag them.

- **Feedback Loop to Users:** To keep users engaged (and producing high-quality votes), show them that their input matters. For instance, after a vote, display how that vote impacted the ratings ("This vote caused Model X's rating to rise by +5.3 Elo points!") or show them a leaderboard of top models to pique interest ("You just helped rank these models. Currently the leader is Y."). LMArena's interface

is straightforward but could be extended in our Arena with more live feedback and maybe fun stats (like “You have voted in 50 debates. You agreed with the AI judge 90% of the time.” etc.).

In essence, **crowd-powered benchmarks thrive on trust and engagement**. By keeping model identities hidden, using Elo to aggregate votes fairly, monitoring for abuse, and making the process enjoyable, we ensure the crowd’s wisdom is both **robust and fun**. Our AI Debate Arena will incorporate these lessons to harness the entertainment value of AI duels while gathering reliable evaluation data.

4. Key Models to Benchmark (November 2025)

The landscape of top-performing LLMs in late 2025 includes several cutting-edge proprietary models and a new generation of open-source models. Our benchmark should include the **most relevant models** in each category:

- **Closed-Source Frontier Models:** These are the flagship models from major AI labs, which often set the state of the art:
- **OpenAI GPT-5.1:** The latest iteration of the GPT series, launched by late 2025. It’s widely regarded as a top performer on general tasks and has likely improved on GPT-4 in areas like reasoning and multimodality. (Indeed, GPT-5.1’s arrival in late 2025 was closely followed by competitors’ releases [46].) We should include GPT-5.1 as a baseline for “best-in-class” performance.
- **Anthropic Claude 4.5 (Claude “Sonnet” 4.x):** Anthropic’s Claude series has evolved (Claude 3, Claude 4, and intermediate versions named after styles like *Opus* and *Sonnet*). By November 2025, Anthropic’s **Claude 4.5** is an advanced conversational model focusing on reliability and longer context. It may not fully match GPT-5 in all areas, but it excels in aligned, human-like responses. Including Claude 4.x (latest version) is important for a well-rounded comparison, as it traditionally has strengths in providing thoughtful, safe answers.
- **Google Gemini 3.0:** Google’s Gemini (the multi-modal successor to PaLM) reached version 3.0 in 2025, and is notable for its **multimodal capacity and massive scale**. Gemini 3 Pro (the top tier) is described as a *trillion-parameter MoE model with up to 1M token context*, excelling at tasks that require vision+text or long-term planning [47] [48]. Early benchmarks show Gemini 3 outperforming GPT-5.1 on certain adversarial reasoning and multimodal tasks [49] [50]. We expect Gemini 3.0 to be a major contender, especially if any debate topics involve interpreting images or diagrams (depending on our platform’s capability).
- **xAI Grok 4.1:** xAI (the Elon Musk-led initiative) introduced **Grok** in late 2023 and by 2025 it’s on version 4.1. Grok 4.1 is positioned as a rival to GPT-4/5 with a unique edge in being an "AI with a bit of attitude" but now more aligned [51]. It reportedly emphasizes creative and emotionally attuned responses [52]. Including Grok 4.1 in the benchmark adds diversity (it may have different training nuances and could excel at persuasive, human-like debate if it indeed was tuned for an “AI companion” feel). According to reports, Grok 4.1 is competitive with other top models, and its presence will be of interest to the community given the high-profile nature of its backer.

- **Other Notable Closed Models:** If available, we might also consider including **Microsoft’s latest GPT-4 iteration** (if they have an in-house variant with enhancements, given MS’s partnership with OpenAI) or **IBM’s Project Debater-derived models** (IBM had work on debate AIs, though not sure if they reached GPT-4 level). Another is **DeepMind’s models** (if DeepMind released a Sparrow or Chinchilla successor by 2025, or if **“Gopher 3”** or others exist). However, the question specifically

names GPT-5.1, Claude 4.5, Gemini 3.0, Grok 4.1, which we've covered. These will likely form the top tier of our Arena.

- **Leading Open-Source Models:** The open-source LLM ecosystem in 2025 has produced models that, while generally a step behind the proprietary giants, are rapidly closing the gap. Some of the top open models to include:

- **Meta's LLaMA 3 and LLaMA 4 series:** Meta continued its open model releases. By 2025, **LLaMA 3.1** and even **LLaMA 4** might be available (with LLaMA 4 rumored to use techniques for extended context and improved reasoning). The HuggingFace Open-LLM list (Nov 2025) indeed cites *Llama 4 (Scout/Maverick)* as a notable model [53]. These models (especially 70B+ parameter versions) are strong baseline chat models and often form the foundation of many fine-tunes.
- **Qwen-3 (235B MoE) by Alibaba: Qwen** (a series from Alibaba) has a massive 235B-parameter mixture-of-experts version (active ~22B per token) that ranks among the best open models [53] [54]. Qwen-3 is known for strong multilingual ability and reasoning, and it has an Apache 2.0 license making it truly open. We should include Qwen-3 as one of the largest open models.
- **"Mixtral" 8×22B Ensemble:** This appears to be a top performer which likely refers to a **Mistral-based ensemble** (Mistral is a new startup model known for efficiency). The entry "Mixtral 8x22B" suggests an ensemble of 8 Mistral 22B expert models [53]. This kind of architecture shows how open models are combining techniques to punch above their weight. Mixtral would be a candidate if publicly available, since it reportedly excels at general chat and reasoning with a large combined parameter count.
- **DeepSeek (V3 R1)**: The mention of **DeepSeek-V3 (R1)** in open rankings [55] implies a powerful open model or series (perhaps from an academic consortium or smaller lab) that has made a mark. DeepSeek might be an open project focusing on very large contexts or robust reasoning. If it's open-source (license to be checked), including it would add a novel entrant.
- **Other notables: Falcon 180B** (from TII UAE) was a top open model in 2023; by 2025 **Falcon 2** or similar might be around, possibly at the top of some leaderboards [56]. **Mistral Large 2** (maybe an updated 70B+ Mistral model) is also cited as a leading open model [57]. **WizardCoder/WizardLM**, **Vicuna 2**, **PaLM 2-based open variants** (if any became open) could also be considered if they bring unique strengths (e.g. a model fine-tuned heavily for coding or a model specialized in knowledge). Given the plethora, we should prioritize those that rank highest on aggregate benchmarks. According to an updated 2025 list, models like *Llama 4, Qwen3, DeepSeek R1, Grok-1 (open)*, **Command R+ (104B)** and **Gemma-2 (27B)** are among the top 10 open models [53] [54]. We'll focus on a representative subset of these to avoid overcrowding.

Including **older model versions** in the benchmark can indeed be useful for **performance tracking over time**. Many leaderboards do retain older models for reference – for example, Chatbot Arena initially included GPT-3.5 and even older baseline bots to show the stark quality differences. We should include a few legacy models such as **OpenAI GPT-3.5 Turbo (2023)**, **Anthropic Claude 2**, and perhaps a classic open model like **GPT-J or GPT-NeoX**. These will likely sit at the lower end of the leaderboard, but they serve as **baseline anchors** and help illustrate progress. Seeing an older model consistently lose debates to newer ones is a tangible demonstration of advancement. It also helps calibrate the difficulty of our debates – if even old models score decently, maybe the topics are too easy; if they score near zero wins, it confirms the newer models are far superior as expected.

Moreover, having historical data (e.g., GPT-4.0 vs GPT-4.2 vs GPT-5.1 over time) is valuable to researchers. Our Arena could become not just a snapshot leaderboard but a timeline of improvement. We might even consider a **"hall of fame"** section: showing how the top score has increased from 2023 to 2025.

To summarize, the Arena should host **all major players**: the latest from the big labs (GPT-5.1, Claude 4.5, Gemini 3, Grok 4.1), the best open models (LLaMA 4, Qwen3, etc.), and a few **reference older models**. This comprehensive roster ensures the benchmark is relevant to both industry leaders and open-source community, while providing continuity with past performance.

5. Designing a Dual Scoring System: User Votes + LLM-Judged Metrics

A core innovation for the AI Debate Arena will be a **dual scoring system** that captures two perspectives on performance: 1. **Crowd Preference Score (Entertainment Metric)** – derived from user votes (the Elo-style ranking as in Chatbot Arena), reflecting which model wins debates in the eyes of human participants. 2. **LLM-Evaluated Score (Quality Metric)** – an automatic scoring of debate transcripts by AI judges, assessing criteria like factual accuracy, logical soundness, and rhetorical quality.

This dual system marries the **engagement of human voting** with the **consistency of AI evaluation**, providing a more rounded benchmark. Key design considerations for this system:

- **Crowd Vote Score (Elo Rating):** We will implement an Elo or Glicko rating system that updates as models win or lose debates against each other. Each debate between two models is essentially a "match." If Model A wins (by human vote), its rating increases and Model B's decreases, with the magnitude of change depending on their prior ratings (so upsets yield bigger changes). Over many battles, this yields a **rank ordering by Elo**, which is intuitive for users (higher Elo = stronger in human eyes). We'll display this as the **"User Vote Rank."** Like LMSYS, we should display a confidence or uncertainty range for each Elo to caution users that close differences might not be significant [8] [16]. The Elo score primarily captures overall **persuasive power and general capability** – if a model consistently wins, it means humans found its arguments better or more helpful across varied topics.

- **LLM-Judged Metrics:** In parallel, after each debate (or periodically on a set of standardized debate prompts), we will employ one or multiple **AI judge models** to analyze the debate transcript. This could involve:

- **Factual Accuracy Check:** An LLM (or a tool-augmented LLM that can consult a knowledge base) evaluates each factual claim made by both sides. It could give a score or just flag major incorrect statements. For instance, a prompt to the judge model: *"Analyze the above debate. Identify any factual inaccuracies from either side, and judge which side was more factually correct overall."* The output can be distilled into a **factual score** (e.g. 8/10 if mostly correct) or a simple win/loss if one side clearly had more truth.
- **Logical Consistency:** The judge looks for logical fallacies or contradictions. It can be asked to rate the coherence of arguments: *Did the debater follow logically or did they stray into non-sequiturs?* This yields a **logic score** or a qualitative label.
- **Rhetorical Quality/Persuasiveness:** Even though human votes reflect persuasiveness, an AI judge can attempt a more **nuanced rhetorical analysis** – evaluating clarity of points, use of strong

evidence, emotional appeal, etc. In structured debate judging (like in human debate tournaments), judges often give points in categories (content, style, strategy). We can mirror a bit of that: the LLM judge could output something like: *"Speaker A: Content 7/10, Style 8/10; Speaker B: Content 6/10, Style 6/10"*, and a verdict of who *should* win on merit. This is subjective for the AI too, but if we use a very advanced, instruction-tuned model for judging (perhaps even GPT-5.1 itself or a specialized judge model), it can provide a consistent rubric-based assessment. Research has found that GPT-4-level judges, while not perfect, often rank outputs similarly to humans in debates [58] [59] – sometimes even finding that top LLMs can craft arguments that *LLM judges* rate higher than human debaters' texts in controlled tests [60].

- **Overall Debate Outcome (AI's perspective):** The AI judge can also simply declare a winner of the debate, independent of the human vote. This effectively simulates a **"referee vote."** We could use this to compute an **AI-Elo rating** by treating the AI's decision as another game result. For example, run each debate transcript through the judge; if the judge disagrees with the human, that's interesting data (but for rating, we might just count it as well, or track separate Elo from AI-judged outcomes).

To implement this, we might utilize multiple models: e.g. GPT-4.1 as a primary judge with a refined prompt, or even train a smaller model on a set of human-judged debate examples to mimic human evaluation. There is emerging work on such *agent-as-a-judge* setups and their reliability [61] [62]. Initially, using a known strong model (like GPT-4 or GPT-5) with a carefully crafted prompt for evaluation is the straightforward path.

- **Combining or Presenting Dual Scores:** We will likely maintain two separate leaderboards or a unified leaderboard with two columns:
- **"Human Elo"** – ranking by crowd preferences.
- **"AI Quality Score"** – which could be a composite of the AI-judged metrics (perhaps normalized to a 0–100 scale or another Elo). For example, a combined score could weight logic, factual, and style categories. Alternatively, we list each metric separately. However, for simplicity and public facing, a single AI score that aggregates those dimensions might be best. We must be transparent about how this score is derived (e.g. "This is the average of logical, factual, and rhetorical scores given by the AI judges across many debates.").

An interesting design is to allow sorting the leaderboard by either metric. A user might toggle between "By Human Votes" and "By AI Evaluation". Ideally, the rankings shouldn't be wildly different (that would indicate a problem either with the human voting process or the AI metrics). But some divergence is expected – e.g. a model that is very persuasive but sometimes hallucinates might rank high in Human Elo but lower in AI score due to factuality penalties. Exposing that gap is valuable: it identifies **"charismatic liars"** vs **"factually reliable"** models. Users could choose based on their preference (for entertainment, they might care who wins human votes; for serious applications, they might care about factual consistency).

- **Calibration and Fairness of AI Judging:** We must ensure our AI judge is not biased towards a particular model or style. Steps to ensure this:
- Use a **neutral judging prompt** that doesn't mention model identities and focuses only on the content of the debate.
- Possibly use a **variety of judges** (e.g. GPT-5 and Claude 4 as two judges) and average their scores, to avoid single-model bias. If they largely agree, that increases confidence.
- Continually validate the AI judge's outputs against a small set of human expert judgments. For instance, occasionally have a human (or ourselves) evaluate a debate for factual accuracy and see if the AI judge spotted the same issues.

- Address known biases: LLM judges can be susceptible to verbosity (thinking a longer answer is better) or formatting (one model might use more structured lists and the judge might favor that systematically). We can try to mitigate this by randomizing formats or explicitly instructing the judge to focus on substance over style formatting. This ties into LMSYS's concept of **style control** to reduce human bias [36] – similarly, we can control style in what the judge sees or how it's prompted.

- **User Visibility and Engagement with Metrics:** To keep the platform fun, we might not show the AI's scoring breakdown to general users (to avoid confusion or making things overly complex during voting). However, on the results or model profile page, we can show something like: *"Model X: Human Elo #3 (Elo 1500), AI Quality Score 85/100 (rank #5)."* For those interested, clicking on Model X could reveal *"AI judge says: Factual 9/10, Logic 8/10, Style 7/10 on average."* This provides an **explanation for its performance** [63]. It also reassures that a model ranked lower by humans might still be very logical, just maybe less "likable" in style, etc.

- **Dual Score Use Case:** Suppose Model A has an Elo rank of 1 (people love it) but the AI score finds it has a habit of making stuff up (factual 5/10). Meanwhile Model B has Elo rank 3 but near perfect factual score. This info would be actionable to users or developers: Model A might be great for creative or opinionated tasks (where human-like flair wins) while Model B might be better for applications needing accuracy. The **epistemic value** of the benchmark is thus enhanced – we're not just crowning the most popular AI, but analyzing the *why* behind wins.

In implementing the dual system, we essentially create a more **engaging yet rigorous evaluation**. The user vote provides the **entertainment factor and direct human preference signal**, and the LLM judge provides a **consistent analytical backbone** across debates. This mirrors how a sports competition might have fan voting for MVP (popular choice) and expert judges scoring performance (technical evaluation) – both are interesting, and comparing them yields insights.

6. Proposal: Upgraded AI Debate Arena Structure and Features

Bringing it all together, we propose an updated **AI Debate Arena** that operates as a credible benchmark while retaining the excitement of AI-vs-AI debates. Below is a structured plan covering data, mechanics, leaderboard design, and evaluation methodology:

Architecture & Data Pipeline:

- **Debate Sessions:** Each debate will be a structured session with a specific **topic or question**, two LLMs (each possibly assigned a *persona/role* relevant to the topic), and a sequence of alternating messages (e.g., Opening statements, Rebuttals, and Conclusions). We can enforce a format (to keep debates comparable) – for example: *Model A opens, Model B opens, Model A rebuts, Model B rebuts, then each gives a closing statement*. Keeping debates to a fixed number of turns ensures fairness and makes evaluation easier, though we might allow longer free-form exchanges in "sandbox" mode for fun.
- **Persona-Driven Setup:** Before the debate, the system will assign or let the user choose **personas** for each model (e.g. "You are a climate scientist defending the reality of climate change" vs "You are a skeptic arguing it's exaggerated"). Personas add flavor and ensure the models have some guidance on style/tone. These persona instructions should be part of the prompt given to the models, and

they remain consistent through the debate. We will store these persona descriptions along with the transcripts as part of the metadata.

- **Data Storage:** All debates (prompt, model outputs each turn, votes, etc.) are stored in a database. Key elements to record:
- Model identifiers and versions involved.
- The prompt/topic and persona instructions.
- Full transcript of the debate.
- The user's vote (who won, or draw if that's allowed).
- The AI judge's evaluation output for that debate (scores or chosen winner).
- Timestamp, plus any user ID (if logged in) or session ID for tracking.

For scalability and analysis, a structured format like JSON lines or a relational schema will be used. This data will feed both the live leaderboard calculations and offline analysis for research. We plan to periodically **export anonymized debate data** (excluding user identity) to share with the community, similar to LMSYS releasing conversation datasets [45]. This supports open research and allows external validation of the benchmark.

- **Real-Time Rating Update:** After each vote, the Elo ratings of the two models are updated immediately in the database. We'll use a standard Elo formula with a suitable K-factor (possibly decreasing as models have more matches to stabilize ratings). We might require a model to accumulate a minimum number of debates (e.g. 20) before its rating is considered "public" on the leaderboard, to avoid volatile ranks from small sample sizes [20] [21]. Until that threshold, it can be listed as "provisional".
- **AI Judge Evaluation:** This can run asynchronously. Right after a debate concludes, we queue the debate transcript for AI evaluation. The judge model (running on a server) processes it and stores the results (could be immediate if using an API or slightly delayed if load is high). These results then update the **AI metrics** for the models. We might aggregate the AI scores over multiple debates for stability: e.g., keep a running average of "factuality score" for each model across its last N debates, updated incrementally. Alternatively, for consistency, we could have a fixed evaluation set: say, every model, upon entry or update, is run through a *gauntlet* of 10 standardized debate scenarios with an AI judge scoring them. This would produce a stable "AI Quality Score" not dependent on the live user-driven debates (which can vary in topic). Or we could do both live and static evaluations. For now, a simpler approach: use the actual user-driven debates themselves as the data for AI scoring, averaged out. This ties the AI evaluation to the same conditions as the human evaluation.

Leaderboard & UI Design:

- **Leaderboard Display:** The main page will feature a table of models with columns for:
- *Model Name & Version*, perhaps with a tooltip or link to more details (like parameter count, source).
- *User Vote Elo* (with rank number and confidence interval). Possibly an icon or color to denote if a model is trending up or down.
- *AI Quality Score* (as a percentile or score out of 100, for example). We can also rank by this score. If needed, multiple columns for sub-scores: e.g. "Logic ", "Factual 📚", "Rhetoric 🎭" with individual ratings, and an overall.
- Perhaps a *Games Played* count (number of debates fought), so users know which ratings are based on lots of data vs new entrants.

The leaderboard might default sort by Human Elo (as that's the "main event"), but with a toggle to sort by AI score or any other column. For casual visitors, this immediately shows two perspectives on "the best AI": e.g. Model X is #1 by humans, but #3 by the AI metric, etc. This design aligns with what we see in some gaming leaderboards where you can sort by different stats.

- **Model Profile Pages:** Clicking a model lets one see detailed stats: win–loss record against each other model (like a head-to-head matrix), average judge scores, and example debates. We might show a chart of its Elo over time (especially if a new version replaced an old, etc.). This page helps with **explainability** – if someone is curious *why* a model is ranked a certain way, they can inspect its transcripts or see that "it tends to lose against Model Y on technical topics," etc. Providing these insights addresses the *score explainability* concern (users shouldn't just see a number, they should be able to trace back to evidence) [63] [26].

- **Visual Design:** To make it entertaining, the site could have **arena theming** – e.g. avatars for models or persona icons in a debate. But importantly for a scientific benchmark, keep it clean and not overly gimmicky. Confidence intervals or error margins should be visible (perhaps as ± values next to Elo) [8]. Perhaps small icons indicate if a model is open-source, what size it is (like "70B" tag), etc., to help contextualize entries.

We might also include filters on the leaderboard: view only open-source models, or only models of a certain size, etc., similar to how Hugging Face allows filtering by model type [64]. This can help users compare within their interest (some may want to find "best open model I can run locally" versus "absolute best overall").

- **Dual Score Visualization:** We could incorporate a simple **scatter plot or radar chart** on the site: e.g. each model as a point with Human Elo on one axis and AI score on another. This can visually show outliers (models that humans love but AI judges rate poorly, and vice versa). It's an additional tool for an "epistemologically meaningful" view – ideally, the best models cluster in the top-right (both humans and AI judge agree on their excellence). If some models lie far off the diagonal, those are interesting cases to investigate. We might embed such a plot on the site (perhaps using a library to update it live as data changes).

Rating Mechanics & Updates:

- **Elo System Details:** We will likely use a Bayesian Elo (like TrueSkill or Glicko) to account for uncertainty which can be reflected as the confidence interval [8]. New models start with a baseline rating (maybe 1500) and high uncertainty; as they get votes, uncertainty narrows. The system should handle draws if we allow them (though typically, Arena doesn't use draws – one winner is chosen by the voter, or we could allow a "tie" vote but that's rare in practice).
- **Periodic Resets or Seasons:** One challenge is when models significantly improve (via retraining) or new versions replace old ones. We will need a policy: if the developer of Model X releases *Model X v2*, do we start a new entry from scratch, or update the existing one? Likely treat it as a new entry (so both appear) to maintain integrity of the historical record, with perhaps an annotation that v2 supersedes v1. The leaderboard can then have multiple lines for "Model X (2025 edition)" vs "(2024 edition)" etc. We could eventually retire old versions from the main view (keeping them in an archive view) if they are no longer relevant, to declutter.

- **Continuous vs Batch Evaluation:** The system will run continuously as users interact. We should also consider running *scheduled evaluation rounds* for the AI metrics: for example, once a day, recalc each model's AI judge score on recent debates or on a fixed set. This ensures consistency day to day (since if one model had a string of easy debates one day, its factual average might temporarily look high, etc.). A nightly batch that re-averages or re-evaluates a standard test set for each model could smooth out noise.

Ensuring Scientific & Epistemic Value:

To ensure the benchmark yields **epistemologically meaningful evaluation** (i.e. it genuinely informs us about models' knowledge and reasoning, not just their ability to play to the crowd), we incorporate the following: - **Diverse and challenging debate topics:** We will curate a set of debate questions that test knowledge and reasoning, not just opinion. For example, topics like *"Resolved: The quantum superposition principle has practical macroscopic effects"* require factual knowledge, whereas *"Should society adopt universal basic income?"* tests reasoning and ethical argumentation. Including a range ensures models must draw on facts and logic, not just style. We can tag debate prompts by category (science, politics, ethics, etc.) and later analyze performance by category – echoing HELM's scenario approach for breadth [32]. - **Persona constraints adding depth:** By giving models specific personas (some with expertise, some with biases), we examine how well models can *inhabit a role and argue consistently*. This tests a form of prompt-following and perspective-taking. A truly capable model should be able to argue either side of an issue with strong points (even if it "believes" one side more). - **Fact-Checking Mechanism:** Perhaps incorporate a phase in debates where models can call for **evidence or sources** (we could allow each model one "fact-check request" per debate, which triggers an external search or a knowledge retrieval that both can see). This would make debates more grounded. If that's too complex to implement initially, we rely on the AI judge to do post-hoc fact checking. - **Evaluation of Failure Modes:** The LLM judge can be instructed to not only score but also provide a critique of each model's performance (e.g. "Model A used a strawman argument; Model B failed to answer a key challenge."). Aggregating these critiques might uncover common failure modes of models. For instance, maybe Model Y often refuses to continue the debate due to a safety guard trigger – that's valuable info. We could present some of this analysis in our research reporting (not necessarily on the public leaderboard, but in a whitepaper or blog). - **User Education:** To balance entertainment with truth-seeking, we might add a small note to voters like "Which argument was more convincing *and sound*?" to subtly encourage quality-based voting, not just superficial preference. If crowd workers or volunteers are particularly quality-focused, the human votes themselves become more epistemic. Crowd moderation (like allowing experienced users to flag if a model said something blatantly false) could also feed into the scoring.

- **Leaderboard Interpretability:** We will include documentation on how to read the leaderboard – emphasizing that the **User Vote rank** reflects subjective preference (which correlates with overall model capability but can be influenced by style), whereas the **AI Quality score** reflects objective criteria. This dual presentation is novel, so users (and even AI developers submitting models) should understand that a "top" model ideally does well on both. If a model only does well on one axis, it might be specialized or have trade-offs.

Finally, to solidify the **Arena's legitimacy as a benchmark**, we plan to: - Publish a methodology report (similar to LMSYS's arXiv paper [22]) detailing our statistical methods, the Elo system, the AI judge calibration, and some initial findings. - Possibly integrate with existing leaderboards: e.g., contribute our results to HELM or have a category on Hugging Face (if they consume external evals). This cross-pollination (like submitting our conversation win-rate data as another metric for models) can raise confidence. The

skywork.ai review recommended treating Arena as one leg of a multi-legged evaluation approach [65] – our goal is to make this leg as strong as possible by internally already combining multiple evaluation aspects.

Conclusion: The upgraded AI Debate Arena will function not just as a playful showdown platform, but as a rich **benchmarking arena**. By incorporating modern evaluation metrics, robust leaderboard practices, crowd-voting safeguards, the latest models, and dual human/AI scoring, it will provide both entertainment **and** reliable insights. Users will be able to enjoy AI personas clashing in debates while the system rigorously tracks who truly has the best arguments. In doing so, AI Debate Arena can help demystify which models are "all talk" and which can back their words with facts and reason – a meaningful step forward in evaluating AI through dialogue.

Sources:

- LMSYS Chatbot Arena methodology and review [8] [14] [22]
- Hugging Face Open LLM Leaderboard description [66] [6]
- Stanford HELM evaluation approach [33] [32]
- LMArena crowd voting and anti-bias policies [13] [67]
- Debates and LLM-as-judge research [3] [58]
- Notable 2025 models and benchmarks [46] [53]

---

1 2 5 6 7 27 28 29 34 66 About
https://huggingface.co/docs/leaderboards/en/open_llm_leaderboard/about

3 4 12 58 59 60 Debatable Intelligence: Benchmarking LLM Judges via Debate Speech Evaluation
https://arxiv.org/html/2506.05062v1

8 14 16 17 18 19 22 23 25 26 30 31 32 33 35 40 43 44 63 65 67 Chatbot Arena (LMSYS) Review 2025: Is the LLM Leaderboard Reliable?
https://skywork.ai/blog/chatbot-arena-lmsys-review-2025/

9 13 20 21 37 38 39 41 42 45 LMArena
https://lmarena.ai/faq

10 Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena - arXiv
https://arxiv.org/abs/2306.05685

11 Chatbot Arena Leaderboard Week 8: Introducing MT-Bench and ...
https://lmsys.org/blog/2023-06-22-leaderboard/

15 What's going on with the Open LLM Leaderboard? - Hugging Face
https://huggingface.co/blog/open-llm-leaderboard-mmlu

24 Demystifying LLM Leaderboards: What You Need to Know - Shakudo
https://www.shakudo.io/blog/demystifying-llm-leaderboards-what-you-need-to-know

36 In the Arena: How LMSys changed LLM Benchmarking Forever
https://www.latent.space/p/lmarena

46 47 48 49 50 Gemini 3 Pro: First Reviews. Launched on November 18, 2025 and... | by Barnacle Goose | Nov, 2025 | Medium
https://medium.com/@leucopsis/gemini-3-pro-first-reviews-527120cebe84

51 52 Grok 4.1 has arrived — and it's bringing the fight to ChatGPT with ...
https://www.tomsguide.com/ai/grok-4-1-has-arrived-heres-what-xais-latest-update-unlocks

53 54 55 10 Best Open-Source LLM Models (2025 Updated): Llama 4, Qwen 3 and DeepSeek R1
https://huggingface.co/blog/daya-shankar/open-source-llms

56 Top 40 Large Language Models (LLMs) In 2025: The Definitive Guide
https://bestarion.com/us/top-large-language-models-llms/

57 Best Open Source LLMs in 2025: Top Models for AI Innovation
https://www.trgdatacenters.com/resource/best-open-source-llms-2025/

61 [PDF] The Rise of Agent-as-a-Judge Evaluation for LLMs - arXiv
https://www.arxiv.org/pdf/2508.02994

62 LLM-as-a-Judge Evaluation - Emergent Mind
https://www.emergentmind.com/topics/llm-as-a-judge-evaluations

64 Understanding LLM Leaderboards: metrics, benchmarks, and why ...
https://toloka.ai/blog/llm-leaderboard/