LIVEANYROLEBENCH: TOWARDS ROLE-PLAYING AS ANY ENTITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly used to simulate diverse roles, from fictional characters in games to real-world entities in social simulations. Yet existing benchmarks largely focus on few well-known characters (e.g., fictional protagonists or historical figures), who already appear in pretraining data, so even high scores may be contaminated by memorization instead of reflecting general simulation ability. Moreover, evaluating only on human characters can yield biased results, leaving performance consistency across fundamentally different entity types (e.g., warships, space missions, or animals) unclear. To address both gaps, we build role-playing tasks directly from Wikipedia, which is a continually updated source that covers an enormous range of entities. By converting each entity article into a sequence of dated actions, we introduce LiveAnyRoleBench, which evaluates whether an LLM can role-play as any entity by predicting its next action given its own history actions as context. This is, to our knowledge, the first role-playing benchmark that dates every task and spans arbitrary entity types (137 types, 25,208 entities, 969,088 actions). Evaluating 13 models, we find that every model's score drops by 3 to 18 percentage points on actions dated after its training cutoff. Across entity types, the strongest model leads almost everywhere, but among models of similar strength, rankings on human characters do not predict rankings on other types. Grounded in real-world actions, LiveAnyRoleBench provides a foundation for role-playing agents and world simulators requiring fine-grained interaction and complex decision-making.[1]

1 INTRODUCTION

Large language models (LLMs) are increasingly used to role-play diverse entities, from fictional characters in games to real people and organizations in social simulations. Role-playing agents portray fictional characters and historical figures in chatbots and video games (Shao et al., 2023; Wang et al., 2025b;a). Social simulations use LLMs to stand in for real people: agents built from interviews with 1,052 people predict those people's survey answers (Park et al., 2024), and SocioVerse draws its agents from a pool of 10 million real users (Zhang et al., 2025). Such simulations can also contain firms, banks, and governments (Piao et al., 2026). For a real entity, what it did next is a fact, and a model helps a simulation only as far as it predicts that fact, for every kind of entity involved.

However, existing role-playing benchmarks mostly test well-known human characters, such as fictional protagonists and historical figures, and this limits what their scores show in two ways. First, the stories of these characters are already in the pretraining data, so a high score may reflect memorization rather than role-playing ability; hiding the characters' names, for instance, lowers role-playing scores (Peng & Chen, 2026). Other evaluations separate memorization from ability by testing on items dated after a model's training cutoff (Jain et al., 2025; White et al., 2025), but role-playing benchmarks do not date their tasks (Table 1). Second, evaluating on human characters alone can bias the comparison of models: CharacterEval's ranking of models agrees with human judgments at Spearman $\rho = 0.19$ on fictional characters but $\rho = -0.34$ on other character categories (Zhou et al., 2025a, Table 6), and how well models role-play companies, warships, space missions, or animals remains largely untested. Addressing both limitations requires tasks that are dated, so that each can be placed before or after a model's training cutoff, and that span every kind of entity.

[1] We release the dataset, construction pipeline, and all model outputs.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

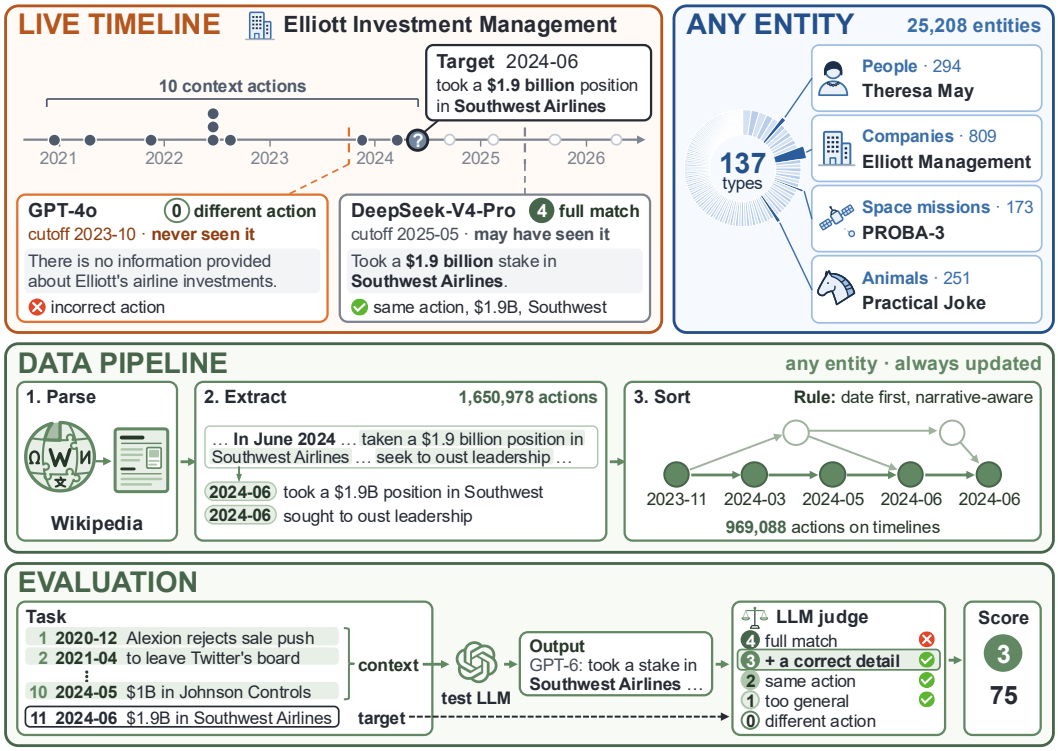


Figure 1: **LiveAnyRoleBench.** *Live timeline*: every action is dated, so each task lies before or after a model's training cutoff; here the target lies after GPT-4o's cutoff, and GPT-4o predicts a different action, and before DeepSeek-V4-Pro's, which predicts it in full. *Any entity*: tasks come from the Wikipedia articles of 25,208 entities of 137 types. *Data pipeline*: an LLM extracts dated actions from each article, and a date-first, narrative-aware sort orders them into a timeline. *Evaluation*: the tested LLM predicts the action that follows ten context actions, and an LLM judge grades it from 0 to 4 (score $= 25\times$ grade).

Our key idea is to build role-playing tasks from Wikipedia: we turn each entity's article into a timeline of dated actions and ask a model to predict what the entity did next. Wikipedia gives these tasks what existing benchmarks lack. Because it dates what entities did, each task can be placed before or after any model's training cutoff, and because it is continuously updated, rerunning the construction on a later dump yields tasks after the cutoffs of newly released models. Because it describes entities of almost every type, from people and fictional characters to companies, football clubs, legislatures, destroyers, space missions, and individual animals, one task format serves them all. And because each answer is what the entity actually did, every prediction is scored against ground truth, which role-play evaluation usually lacks (Zhou et al., 2025b), and building the tasks needs no hand-built environment, no interacting agents, and no human annotation of targets.

Turning an article into tasks requires a dedicated pipeline, because an article is not a timeline of decisions: it is ordered by topic rather than by time, most of its dates are coarse, and many of its sentences report what happened to the entity rather than what the entity decided. In our pipeline, an LLM extracts dated actions sentence by sentence; a date-first, narrative-aware sort orders them into a timeline, using the article's order within a section where the dates cannot decide; and only actions the entity could have planned become targets. On 25,208 entities, the pipeline yields 969,088 ordered actions for US$1,175 at list prices.

The resulting benchmark, **LiveAnyRoleBench**, evaluates whether an LLM can role-play as any entity by predicting its next action from its own history of actions. Each task gives a model ten consecutive dated actions of an entity and a question that names who acts and what the next action concerns without revealing it; an LLM judge scores the prediction against the action that actually followed, the target, from 0 to 100. Because every task is dated, it is labelled, for each model, as

before or after that model's training cutoff, so one benchmark separates what a model may have read from what it must infer. In Figure 1, Elliott Investment Management's $1.9 billion position in Southwest Airlines (June 2024) comes after GPT-4o's cutoff, and GPT-4o predicts a different action; DeepSeek-V4-Pro, whose cutoff is later, may have read about it and predicts it in full. Because all 137 entity types share one task, models can be ranked on companies, warships, or animals as well as on people. To our knowledge, LiveAnyRoleBench is the first role-playing benchmark that dates every task and spans arbitrary entity types (Table 1). Grounded in real-world actions, it provides a foundation for role-playing agents and world simulators that require fine-grained interaction and complex decision-making.

We evaluate 13 LLMs, released between July 2024 and September 2026, on 20,109 tasks from 1,000 entities of all 137 types, and find that both limitations matter. Every model's score drops by 3 to 18 percentage points on actions dated after its training cutoff, and the strongest model, GPT-6 Astra, drops the most (Figure 3). The drop is not because newer actions are harder: the four models with training cutoffs in 2023 cannot have seen any action from 2024 to 2026, and they score no lower on the 2025 and 2026 actions than on the 2024 ones. Scores on actions a model may have read about thus overstate its role-playing ability. Across entity types, GPT-6 Astra is best on 72 of 83 types, but below it the order changes: four models that tie on human characters differ by up to 25 points on other types (Figure 4), so a ranking on human characters alone cannot tell which of them role-plays a company or an animal best.

Our contributions are as follows:

- A construction pipeline that turns Wikipedia articles into dated next-action tasks for any entity type and, rerun on a later dump, extends them to newly released models (§2).
- LiveAnyRoleBench, with 25,208 entities of 137 types, 969,088 ordered actions, and a before/after label for every task at each evaluated model's training cutoff (§2).
- An evaluation of 13 LLMs, showing that every model scores lower after its training cutoff and that, among models of similar strength, rankings on human characters do not predict rankings on other entity types (§3); we release the dataset, the construction pipeline, and all model outputs.

2 LIVEANYROLEBENCH

2.1 BENCHMARK OVERVIEW

As illustrated in Figure 1, we turn each entity's English Wikipedia article into a dated, confidently ordered timeline of actions, mark the actions the entity could have planned, and cut the timeline into next-action prediction tasks: ten consecutive actions form the context, and the action that follows them, which must be plannable, is the target to predict. The full build covers 25,208 entities of 137 entity types, with 969,088 confidently ordered actions and 438,792 tasks (Table 3 in Appendix A).

All experiments use an evaluation split of 1,000 entities and 20,126 tasks, small enough to run every evaluated model on every task. Its 716 *representative* entities are those that yield tasks among 1,000 entities drawn from the 25,000 by the same stratified sampling. Its 284 *fresh* entities come from the same article pool and were chosen for recent activity, because a faithful sample is mostly about the past: only 3.0% of the targets of the 25,000 sampled entities start on or after 18 July 2024, against 11.5% in the split. Table 1 summarizes the differences from prior benchmarks.

2.2 DATA CURATION

Source and entities. We use the English Wikipedia dump of 1 July 2026. A candidate entity is an article whose Wikidata *instance of* (P31) class can hold an identity and act; the pool holds 99,997 articles in 567 classes, at most 312 per class. We draw 25,000 entities by proportional stratification over the classes, sampling systematically along article length within each class so that the sample keeps the size distribution of the pool (median article 16,563 bytes in the pool, 16,566 in the sample). Together with the 208 fresh entities of the evaluation split (§2.1) that are not in this sample, they form the full build of 25,208 entities. The 567 P31 classes are folded into 137 entity types that a reader can recognize by name (Section 2.4), and every task keeps the revision id of its article.

Table 1: **Only LiveAnyRoleBench has all four properties.** Cutoff split: items are labelled before or after each evaluated model's training cutoff; –: no or not applicable.

Benchmark	Live		Any entity			
	Dated items?	Cutoff split?	#Entities	#Types	Non-person?	Next action?
CoSER (Wang et al., 2025b)	–	–	17,966	1	–	–
CharacterBench (Zhou et al., 2025a)	–	–	3,956	25	✓	–
BehaviorChain (Li et al., 2025b)	–	–	1,001	2	–	✓
RoleLLM (Wang et al., 2024a)	–	–	100	1	–	–
PersonaGym (Samuel et al., 2025)	–	–	200	1	–	–
SimBench (Hu et al., 2026)	–	–	–	1	–	–
OPeRA (Wang et al., 2026)	✓	–	51	1	–	✓
LiveCodeBench (Jain et al., 2025)	✓	✓	–	–	–	–
AntiLeakBench (Wu et al., 2025b)	✓	✓	–	–	–	–
Daily Oracle (Dai et al., 2025)	✓	✓	–	–	–	–
LiveAnyRoleBench	✓	✓	**25,208**	**137**	✓	✓

From sentences to actions. We clean the wikitext, split it at headings, cut it into chunks of at most 1,500 characters, and number the sentences. Curation calls a model twice. The first call (`gpt-4.1-mini`) returns, for every sentence, either action rows {`sid, time, action`} or a `state` row for a sentence that reports no change (Vendler, 1957). The index `sid` ties each action to its source sentence, and 98.6% of actions take the form `<Actor>: <verb phrase>`, from which the acting party is read.

A confidently ordered timeline. An article is ordered by topic, and most of its dates are coarse, so sorting actions by date with article order as the tie-break fails: on a reference chronology of three articles (708 dated actions), 57.5% of the 677 eleven-action windows this rule produces contradict the reference, most often because a year-only date sorts ahead of finer dates in the same year. We instead build a partial order. Each date becomes an interval, and article order counts as temporal order only within a section, where it is right on 83% of adjacent pairs in the reference, against 36% across sections. One action precedes another if its interval ends before the other's begins or a within-section path links them; bounds are propagated along this order to a fixpoint, and a within-section link that the dates contradict, or that closes a cycle, is dropped. On the reference articles this lowers the share of contradicted windows from 57.5% to 23.9%. To raise coverage, an undated action takes the interval between its dated neighbors in the same section, and an action dated backwards against its section leaves the section order but keeps its interval. The timeline is the longest path through the graph; on the reference articles it grows from 211 actions under dates alone to 463.

Plannable actions. Many timeline actions were not the entity's decision: an outsider acted on it, an outcome became known, or a state held. The second call (`gpt-4.1`) labels each timeline action *plannable* if the entity, or someone acting for it, decided to do it and could have written it into a plan beforehand, and *unplannable* otherwise; sending a probe to a moon is plannable, what the probe finds there is not. On 90 actions labelled by one of the authors, the labels agree with the reference on 79–81 of them across three runs. In total, the 25,208 entities' 2,131,390 sentences yield 1,650,978 actions; 1,347,825 of them (82%) carry a date interval, 969,088 (59%) lie on a timeline, and 580,201 of those (60%) are plannable. Both model calls run at temperature 0 with every prompt and reply kept; building all 25,208 entities cost US$1,175 at list prices; all other steps are deterministic code, and human effort went only into author-labelled reference sets for validation.

2.3 Task Building

Context and target. A task takes eleven consecutive actions of an entity's timeline: the first ten are the *context*, and the eleventh is the *target*, the ground truth a prediction is scored against. Only the target must be plannable, so every plannable action with at least ten predecessors yields one task, 438,792 in all; 6,848 of the 25,208 entities (27.2%) yield none, 88% of them because their timeline

is shorter than eleven actions. Since actions that cannot be confidently ordered stay off the timeline, the target is the next confidently ordered action, not necessarily the very next thing the entity did.

The question. An entity usually has many plausible next actions, and a prediction that names a different but reasonable one cannot be scored against the single target. Each task therefore carries a question that fixes who acts and what the target is about, written so as not to reveal what happened: `What happened next with <Actor> regarding <area>?` The acting party is the target's own prefix, which names the entity itself in only 52% of evaluation targets; the area is written by `gpt-4.1-mini` (temperature 0) from the target, with the target's names, titles, numbers and dates forbidden. For the target *"Philadelphia Eagles ownership: opted to promote assistant coach Nick Skorich instead of Van Brocklin as head coach"*, the question is *"What happened next with Philadelphia Eagles ownership regarding head coaching succession?"*

Per-model Live labels. Since every action carries a date interval, a task can be labelled against any cutoff date D: it is BEFORE if its target ends before D, AFTER if its target starts on or after D, and STRADDLING if the target's interval contains D; straddling tasks are left out of the Live analysis. D is the model's training cutoff as its vendor states it; for a model whose vendor states none, we use an unofficial estimate and mark it (Table 4). Self-reported cutoffs are unusable: asked directly, 11 of 17 OpenAI models answered exactly "June 2024".

2.4 BENCHMARK PROPERTIES

Entity diversity. The 137 entity types (Table 7, Appendix G) run from *People*, *Companies* and *Universities* through *Legislatures*, *Ministries* and *Armed groups* to *Regiments*, *Submarines*, *Space missions* and *Animals*. No type dominates the full build: every type has at least 291 tasks from at least 40 entities, the largest holds 5.1% of the tasks, and the normalized entropy of tasks over types is 0.93 (at most 1; Figure 2). In the evaluation split, 83 types have at least 5 entities and 50 tasks, enough for per-type results (Section 3). Only 2.7% of the full build's tasks come from people and 2.4% from fictional characters, so to score well a model must predict what football clubs, ministries and warships do, not only people.

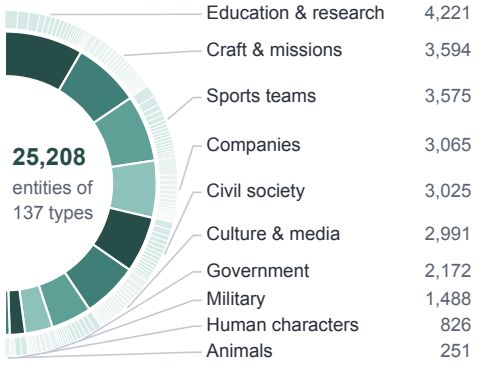


Figure 2: **Entities per domain** in the full build; the outer ring splits each domain into its entity types (Table 7).

Temporal coverage. In the full build, 81.5% of the targets carry a date stated in the article, 45% are dated to the month or finer, and the central 90% fall between 1805 and 2023, so most tasks concern a past that any model may have read about (Table 3). The evaluation split moves the median target year from 1984 to 2001, and it holds 110 to 2,730 after tasks for each of the 13 evaluated models; 76 targets dated after the dump are announced plans and are set aside. An after target had not happened when the model's training data ended, so the model cannot have memorized it and must predict it from the context; a later dump extends the after tasks to newer models.

3 EXPERIMENTS

3.1 SETUP

Models. We evaluate 13 LLMs released between July 2024 and September 2026, so that their training cutoffs fall at different points of the benchmark's timeline: seven proprietary models (GPT-4o (OpenAI, 2024), GPT-4.1-mini (OpenAI, 2025a), GPT-5 (OpenAI, 2025b), GPT-5.5 (OpenAI, 2026a), and GPT-6 Astra, Sol and Luna (OpenAI, 2026c;d)) and six open-weight models (DeepSeek-V4-Pro (DeepSeek-AI, 2026a), DeepSeek-V4.1-Flash (DeepSeek-AI, 2026b), Qwen2.5-72B-Instruct and Qwen2.5-7B-Instruct (Yang et al., 2024), Qwen3-8B (Yang et al., 2025)

Table 2: **Main results by entity domain.** Mean score (0–100, higher is better) of the 13 models on 20,109 tasks from 1,000 entities, overall (**All**) and within each entity domain; every entity counts once. Rows are sorted by All; domains are ordered from easiest to hardest by the mean over the 13 models. **Bold**: highest score in the column. Domain names are shortened; Appendix G lists their entity types, and Table 6 (Appendix F) gives the 95% interval of each All score.

Model	**All**	Sport	Craft	Education	Society	Government	Culture	Military	Companies	Animals	Human
entities	1,000	142	173	139	115	95	130	53	118	14	21
GPT-6 Astra	**55.1**	**56.9**	**58.3**	**56.0**	**55.5**	**54.0**	**51.1**	**51.8**	**52.7**	**60.7**	**54.6**
GPT-5.5	48.0	51.6	50.3	48.6	46.2	46.3	46.2	45.5	45.9	53.0	44.7
GPT-6 Sol	45.6	49.8	46.7	44.9	44.1	47.3	43.6	40.8	44.1	47.5	44.2
DeepSeek-V4-Pro	40.9	44.4	42.5	41.4	40.3	38.3	38.4	40.1	39.1	51.7	37.4
GPT-6 Luna	40.6	43.9	41.9	42.3	40.5	39.4	37.9	36.4	39.4	40.9	36.4
GPT-5	37.9	40.4	38.7	40.2	36.9	39.5	35.3	35.5	36.0	32.7	36.3
GPT-4o	35.9	39.3	36.5	36.8	36.7	34.2	34.7	33.8	34.1	27.0	35.3
DeepSeek-V4.1-Flash	34.4	38.6	36.6	32.4	34.8	33.8	31.6	33.8	31.9	38.4	31.5
GPT-4.1-mini	33.3	38.0	35.6	33.6	33.3	31.6	31.7	32.3	29.7	22.9	29.3
Qwen2.5-72B	28.0	30.4	27.6	28.4	28.0	28.6	27.8	27.9	26.3	17.9	27.2
Qwen3-8B	26.1	28.4	27.6	27.2	25.9	24.8	26.9	24.5	23.1	15.2	21.8
Llama-3.1-8B	25.3	25.1	27.2	26.6	23.2	23.1	27.5	23.7	24.0	26.0	22.8
Qwen2.5-7B	23.9	25.5	25.3	23.9	22.7	25.0	25.1	22.6	21.2	12.2	22.2

and Llama-3.1-8B-Instruct (Grattafiori et al., 2024)). Each model answers each task once, at temperature 0 wherever the API accepts it. All models answer without a reasoning phase except GPT-6 Astra, whose lowest setting (*low*) spends 93 reasoning tokens per task on average. Table 4 (Appendix B) lists identifiers, release dates, training cutoffs and settings. The vendors of four models (DeepSeek-V4.1-Flash and the three Qwen models) state no training cutoff; for these the Live analysis uses a tentative estimate, marked † in every table and figure (Appendix C).

Evaluation split and prompt. All results are on the 1,000-entity evaluation split (§2.1), without the 17 tasks that serve as worked examples in the judge's prompt: 20,109 tasks per model and 261,414 judged predictions in total.[2] Each task is sent as a single user message with no system prompt and no examples: the entity's name and Wikidata class, the ten context actions numbered in order with their dates or date intervals, the task's question, and the instruction *"Answer with a single past-tense sentence describing one action. Do not explain."* (Appendix D).

Scoring. A target is a free-text action, and a prediction can match it at different depths, so we score it on five tiers. **Tier 0**: the prediction repeats one of the last two context actions or describes a different happening. **Tier 1**: it is true of the target but too general to identify it. **Tier 2**: it describes the same happening, but gives no correct detail beyond what the prompt already contains. **Tier 3**: it also gives at least one correct detail of the target—a name, number, title or date—that the prompt does not. **Tier 4**: it covers the target in full, with nothing missing or wrong. Whether two happenings are the same is decided with names set aside: *sold the firm to Buyer X* and *sold the firm to Buyer Y* are the same happening, so a wrong name costs a prediction tiers 3 and 4 but never turns it into a different action. A prediction's score is 25 times its tier, so scores run from 0 to 100, and we report the mean score.

The judge (Zheng et al., 2023) is GPT-5.6 Luna (OpenAI, 2026b) with reasoning disabled. For each task it reads the target, the question, the last two context actions and up to four predictions at once; it analyses the target first, then answers five yes/no questions per prediction, and a fixed rule maps the answers to the tier, so the judge never states a score (Appendix E). On 200 predictions (50 tasks $\times$ 4 of the evaluated models) labelled by one of the authors, the judge gives the author's tier to 162–164 of the 200 across three orderings of the predictions.

3.2 Main Results

Table 2 reports each model's mean score, averaged within each entity and then over entities so that every entity counts once, over all entities and within ten domains that group the 137 entity types.

[2] One judge request, holding three predictions for a single task, failed twice at the output limit and is missing.

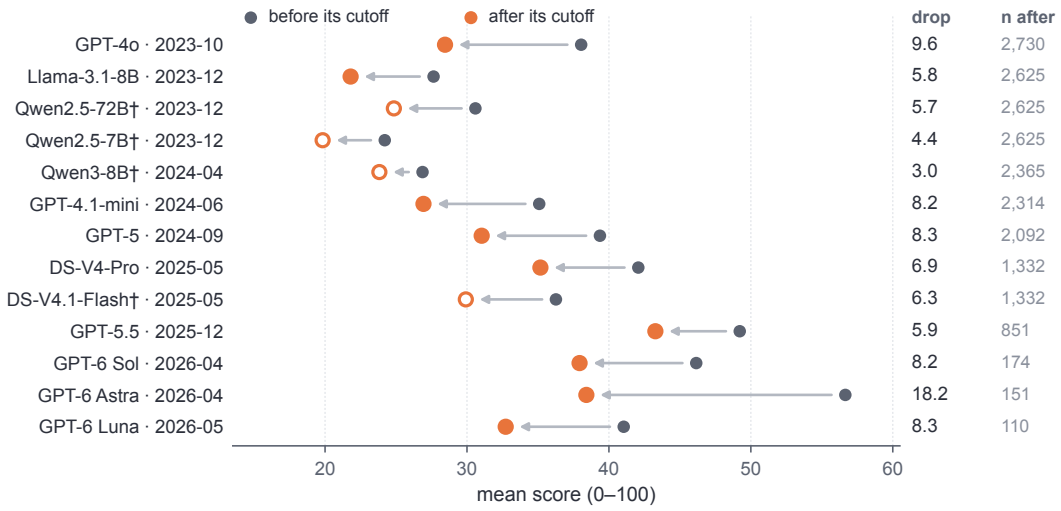


Figure 3: **Every model scores lower on actions after its own training cutoff.** Mean score on tasks whose target lies before vs. after each model's training cutoff; rows are ordered by cutoff, and n after counts the tasks after the cutoff. † Tentative cutoff estimate (Appendix C); hollow markers. DS: DeepSeek.

GPT-6 Astra, the only model that reasons before answering, leads with 55.1, followed by GPT-5.5 (48.0) and GPT-6 Sol (45.6); DeepSeek-V4-Pro (40.9) is the strongest open-weight model, and the three models of at most 8B parameters score between 23.9 and 26.1. The tier distribution shows where the points are lost (Table 6): GPT-6 Astra covers the target in full on 30.1% of tasks and scores 0 on 24.9%, against 6.6% and 57.1% for Qwen2.5-7B. A prediction that is true of the target but too general (tier 1) is rare for every model (2.8–11.2%); the scale mostly separates predicting the wrong action from predicting the right one with or without its particulars.

3.3 Live: Performance Before and After the Cutoff

Every model scores lower after its training cutoff. Figure 3 splits each model's tasks at its training cutoff. All 13 models score lower on tasks whose target lies after the cutoff, by 3.0 (Qwen3-8B) to 18.2 (GPT-6 Astra) points, and every 95% interval excludes zero (Table 8). The strongest model drops the most: GPT-6 Astra scores 56.7 before its cutoff and 38.4 after it. Replacing the tentative cutoffs of the four models without a vendor statement by any other value we tested leaves every drop significant (Appendix C). The Live analysis leaves out the 76 targets dated after the newest article snapshot (July 2026): they are announced plans carried at their future date (*"The film: was further delayed to June 25, 2027"*), not actions that have happened.

Newer actions are not harder in themselves. The actions after a model's cutoff are also the most recent ones, so the drop would also appear if recent actions were harder for every model. We check this on the four models whose training cutoffs fall in 2023 (GPT-4o, Llama-3.1-8B, and the tentative Qwen2.5-72B and Qwen2.5-7B), none of which can have seen an action from 2024 to 2026. On the tasks after the cutoff of all four, their mean score is 22.6 on 2024 actions, 22.3 on 2025 actions and 26.7 on 2026 actions (95% entity-bootstrap intervals 20.5–24.9, 20.4–24.3 and 24.3–29.2). The 2026 rise comes from a different mix of entities: within the 143 entities with actions both in 2024–25 and in 2026, the score changes by $+0.5$ (-2.9 to $+3.9$). The drop after each model's cutoff is therefore not explained by recent actions being harder.

3.4 Any: Performance Across Entity Types

The strongest model leads almost everywhere. Over the 83 entity types with at least five entities and 50 tasks, averaging each model's score within each entity and then over the type's entities, GPT-6 Astra is the best model on 72.

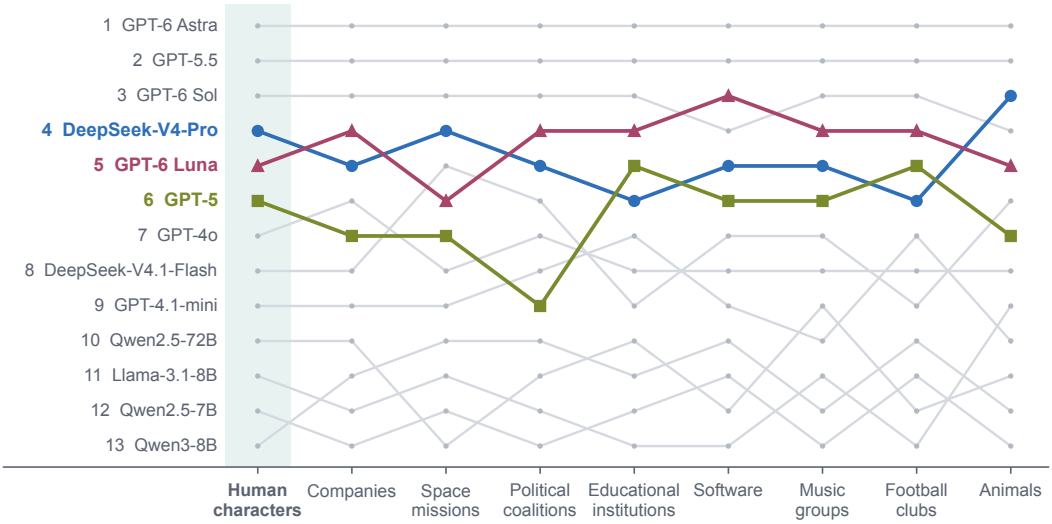


Figure 4: **Among models of similar strength, ranks on human characters do not predict ranks on other entity types.** Rank of each model by mean score on the 21 human characters (shaded) and on the eight largest other entity types in the evaluation split, one per kind; scores are averaged within each entity, then over the type's entities. Coloured: the models ranked 4th to 6th on human characters.

Below it, rankings on human characters do not predict rankings on other types. Figure 4 follows each model's rank from the 21 human characters (people and fictional characters) to the eight largest other entity types of the evaluation split, one per kind. GPT-6 Astra, GPT-5.5 and GPT-6 Sol hold the top three places on almost every type, but the models below them change order. DeepSeek-V4-Pro, GPT-6 Luna, GPT-5 and GPT-4o score 37.4, 36.4, 36.3 and 35.3 on human characters, and every pairwise 95% interval includes zero; on the other eight types, the same four models are up to 24.8 points apart (on animals, DeepSeek-V4-Pro scores 51.7 and GPT-4o 27.0). GPT-5, sixth on human characters, ranks between fifth and ninth on the other types, and DeepSeek-V4-Pro ranks third on animals but sixth on educational institutions and football clubs. An evaluation on human characters alone cannot tell these models apart, although they differ on other entity types.

4 Related Work

Role-playing benchmarks. Role-playing benchmarks mostly test well-known characters from novels, scripts, and games (Wang et al., 2024a; Tu et al., 2024; Wang et al., 2024b; 2025a; Wu et al., 2025a; Ngokpol et al., 2025), 17,966 of them from 771 renowned books in CoSER (Wang et al., 2025b), as well as historical figures (Shao et al., 2023) and characters at a given point of their story (Ahn et al., 2024; Song et al., 2026); PersonaGym instead uses profiles generated by GPT-4o (Samuel et al., 2025). CharacterBench, the broadest, spans 25 character categories and notes that earlier benchmarks "often only involve a single character category" (Zhou et al., 2025a). Because these characters are well documented, scores can reflect memorization: the authors of CoSER "cannot confirm whether a specific LLM has thoroughly learned about a particular book", so comparing pre-trained models "may not be entirely fair" (Wang et al., 2025b), and anonymizing names lowers scores (Peng & Chen, 2026). LiveAnyRoleBench instead draws entities of 137 types from Wikipedia, from people and companies to space missions and animals, and dates every task.

Behavior simulation and next-action prediction. Social simulation places LLM agents in sandboxes and societies (Park et al., 2023; Piao et al., 2026; Zhang et al., 2025), and recent work compares simulated with real survey answers (Park et al., 2024; Hu et al., 2026), Reddit threads (Yu et al., 2026), and individual reasoning (Li et al., 2025a). Closest to our task, OPeRA (Wang et al., 2026) and Lu et al. (2026) predict a shopper's next web action from real sessions, and BehaviorChain the next behavior of characters from fiction and biographies (Li et al., 2025b). Codified

role-playing methods predict a fictional character's next action in a scene and score it against the actual one with an NLI judge (Peng & Shang, 2025; Peng et al., 2026b;a), and GROVE carries this setting to 44 real organized groups whose decisions are mined from Wikipedia and TechCrunch (Zou et al., 2026). Each studies a single kind of subject: people, real or fictional, or organized groups. LiveAnyRoleBench asks the same next-action question of entities from 137 types.

Contamination and live benchmarks. Beyond decontaminated, closed test sets (Zhao et al., 2025), dynamic benchmarks keep adding new items (Chen et al., 2025). LiveCodeBench dates contest problems and scores each model only on problems released after its cutoff, or after its release date when no cutoff is published (Jain et al., 2025); LiveBench adds recent questions every month (White et al., 2025); AntiLeakBench builds questions from Wikidata facts that changed after a model's cutoff (Wu et al., 2025b); and others draw on fast-changing knowledge, new arXiv papers, Wikipedia edits, or events not yet resolved (Vu et al., 2024; Shabtay et al., 2025; Franzmeyer et al., 2024; Karger et al., 2025; Dai et al., 2025). Reported cutoffs can differ from effective ones (Cheng et al., 2024), and post-cutoff drops depend on how the questions are built (Zhang et al., 2026). None of these benchmarks evaluates role-play. In role-play and behavior simulation, time is usually a character's own (Ahn et al., 2024; Han et al., 2026), and cuts by recency are fixed rather than set per model: BehaviorChain holds out recently published books (Li et al., 2025b), and GROVE adds TechCrunch decisions from January 2025 to March 2026, after most models' training cutoffs (Zou et al., 2026). LiveAnyRoleBench dates every task, so each task can be labelled before or after each evaluated model's training cutoff.

5 Conclusion

We introduced LiveAnyRoleBench, which turns Wikipedia's dated accounts of what entities did into next-action prediction tasks across 137 entity types and labels every task before or after each evaluated model's cutoff. On 13 models, every model scores lower on actions dated after its training cutoff. Across entity types, the strongest model leads almost everywhere, but models of similar strength change order from one type to another, so rankings on human characters alone can mislead. Three limitations bound these results. The evaluation split covers 1,000 of the 25,208 entities, so each per-type estimate rests on a few entities. Scores come from one LLM judge, which matches an author's labels on 81–82% of predictions and comes from the same vendor as seven of the evaluated models. And roughly one target in eight, at the plannable labeller's measured error rate, is not an action the entity could have planned. Because Wikipedia keeps growing, running the same pipeline on a later dump extends the after-cutoff tasks to newly released models without changing anything else. We release the dataset, the construction pipeline, and all model outputs.

AI Use Statement

LLMs are components of the benchmark's construction and evaluation. In construction, a local Qwen3-8B screened Wikidata classes for whether their instances can act, `gpt-4.1-mini` extracts dated actions and phrases each question's area, and `gpt-4.1` labels actions as plannable (Sections 2.2 and 2.3); in evaluation, an LLM judge (GPT-5.6 Luna) scores predictions (Section 3.1). The authors reviewed the class list and the 137 types by hand, checked the extracted dates against a reference chronology (itself built with AI assistance), and validated the plannable labels and the judge against an author's labels (79–81 of 90 actions and 162–164 of 200 predictions agree); individual tasks are not checked by hand. Two of the evaluated models, GPT-4.1-mini and Qwen3-8B, also serve as construction tools.

We also used generative AI tools to assist with writing and editing, code for the pipeline and the analyses, literature search, and figures and tables. We did not use generative AI tools for research ideation. The authors checked all AI-assisted text, code, references, figures and tables, and take full responsibility for the content of this work.

Ethics Statement

LiveAnyRoleBench is built from the English Wikipedia dump of 1 July 2026. Wikipedia's text is licensed under CC BY-SA 4.0; the benchmark contains text derived from it and will be released

under the same licence, with every task linked to its source article and revision. The entities are real, and some are living people. We use only what their public Wikipedia articles report, collect no other personal data, and do not contact or interact with any entity. The benchmark scores predictions of actions that have already been reported; it is meant for evaluating LLMs, not for forecasting or profiling particular people or organizations. The reference labels for the judge were produced by one of the authors; no crowdworkers or other human subjects took part. Wikipedia covers some regions, languages and groups more fully than others, and the benchmark inherits this: results for a type describe the entities of that type that Wikipedia documents.

Reproducibility Statement

Section 2.2 describes how entities are drawn and how articles become tasks, and Appendix A gives the resulting statistics. Appendix D gives the prediction prompt and the question template, Appendix E the judge's questions and the mapping from answers to tiers, and Appendix B the exact model identifiers, release dates, training cutoffs and decoding settings. Every task keeps the revision id of its source article, so its source text can be retrieved. We will release the benchmark, the construction and evaluation code, and all model predictions and judge outputs upon publication.

References

Jaewoo Ahn, Taehyun Lee, Junyoung Lim, Jin-Hwa Kim, Sangdoo Yun, Hwaran Lee, and Gunhee Kim. TimeChara: Evaluating point-in-time character hallucination of role-playing large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3291–3325. Association for Computational Linguistics, 2024.

Simin Chen, Yiming Chen, Zexin Li, Yifan Jiang, Zhongwei Wan, Yixin He, Dezhi Ran, Tianle Gu, Haizhou Li, Tao Xie, and Baishakhi Ray. Benchmarking large language models under data contamination: A survey from static to dynamic evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10080–10098. Association for Computational Linguistics, 2025.

Jeffrey Cheng, Marc Marone, Orion Weller, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. Dated data: Tracing knowledge cutoffs in large language models. In *First Conference on Language Modeling*, 2024.

Hui Dai, Ryan Teehan, and Mengye Ren. Are LLMs prescient? A continuous evaluation using daily news as the oracle. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 2025.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026a.

DeepSeek-AI. Introducing DeepSeek-V4.1-Flash: smarter, faster, more efficient. `https://www.deepseek.com/en/news/deepseek-v4-1-flash/`, 2026b. September 10, 2026.

DeepSeek-AI. Public summary of training content for DeepSeek-V4. `https://cdn.deepseek.com/policies/DeepSeek_Template_for_the_Public_Summary_of_Training_Content_for_GeneralPurpose_AI_model.pdf`, 2026c. EU AI Act template; accessed 25 September 2026.

Tim Franzmeyer, Aleksandar Shtedritski, Samuel Albanie, Philip Torr, Joao F. Henriques, and Jakob Foerster. HelloFresh: LLM evalutions on streams of real-world human editorial actions across X community notes and Wikipedia edits. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12702–12716. Association for Computational Linguistics, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Sunkyung Han, Nahyeon Park, Gaeun Seo, Seunghyun Yoon, and JinYeong Bak. CHARM: Character hallucination for multicultural role play benchmark. *arXiv preprint arXiv:2609.01352*, 2026.

Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. SimBench: Benchmarking the ability of large language models to simulate human behaviors. In *The Fourteenth International Conference on Learning Representations*, 2026.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and contamination free evaluation of large language models for code. In *The Thirteenth International Conference on Learning Representations*, 2025.

Haozhe Jia. HindsightBench: A black-box behavioral audit protocol for parametric hindsight in time-indexed LLM decision tasks. *arXiv preprint arXiv:2607.18867*, 2026.

Ezra Karger, Houtan Bastani, Yueh-Han Chen, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E. Tetlock. ForecastBench: A dynamic benchmark of AI forecasting capabilities. In *The Thirteenth International Conference on Learning Representations*, 2025.

Chance Jiajie Li, Zhenze Mo, Yuhan Tang, Ao Qu, Jiayi Wu, Kaiya Ivy Zhao, Yulu Gan, Jie Fan, Jiangbo Yu, Hang Jiang, Paul Pu Liang, Jinhua Zhao, Luis Alberto Alonso Pastor, and Kent Larson. HugAgent: A human simulation benchmark for individual-level reasoning. *arXiv preprint arXiv:2510.15144*, 2025a.

Rui Li, Heming Xia, Xinfeng Yuan, Qingxiu Dong, Lei Sha, Wenjie Li, and Zhifang Sui. How far are LLMs from being our digital twins? A benchmark for persona-based behavior chain simulation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 15738–15763. Association for Computational Linguistics, 2025b.

linux.do. Forum thread on the knowledge cutoff stated by DeepSeek's web model in pre-release testing (in Chinese). `https://linux.do/t/topic/2404493`, 2026. June 2026; accessed 25 September 2026.

Yuxuan Lu, Jing Huang, Yan Han, Bingsheng Yao, Sisong Bei, Yaochen Xie, Yisi Sang, Qi He, and Dakuo Wang. Can LLM agents simulate multi-turn human behavior? Evidence from real online customer behavior data. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 43961–43977. Association for Computational Linguistics, 2026.

models.dev. Model database, entries for Qwen2.5-7B-Instruct and Qwen2.5-72B-Instruct. `https://models.dev`, 2026. Field `knowledge`; accessed 25 September 2026.

Perapard Ngokpol, Kun Kerdthaisong, Pasin Buakhaw, Pitikorn Khlaisamniang, Supasate Vorathammathorn, Piyalitt Ittichaiwong, and Nutchanon Yongsatianchot. Beyond one world: Benchmarking super heros in role-playing across multiversal contexts. *arXiv preprint arXiv:2510.14351*, 2025.

OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

OpenAI. Introducing GPT-4.1 in the API. `https://openai.com/index/gpt-4-1/`, 2025a. April 14, 2025.

OpenAI. OpenAI GPT-5 system card. *arXiv preprint arXiv:2601.03267*, 2025b.

OpenAI. GPT-5.5 system card. `https://deploymentsafety.openai.com/gpt-5-5`, 2026a. April 23, 2026.

OpenAI. GPT-5.6 preview system card. `https://deploymentsafety.openai.com/gpt-5-6-preview`, 2026b. June 26, 2026; covers GPT-5.6 Sol, Terra, and Luna.

OpenAI. GPT-6 Astra system card. `https://deploymentsafety.openai.com/gpt-6-astra`, 2026c. September 3, 2026; appendix on GPT-6 Sol and GPT-6 Luna added September 22, 2026.

OpenAI. GPT-6 Luna model. https://developers.openai.com/api/docs/models/gpt-6-luna, 2026d. Model page with the training cutoff (May 18, 2026); accessed September 24, 2026.

OpenRouter. Models API, field knowledge_cutoff. https://openrouter.ai/api/v1/models, 2026. Accessed 25 September 2026.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22. ACM, 2023.

Joon Sung Park, Carolyn Q. Zou, Jonne Kamphorst, Niles Egan, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Percy Liang, Robb Willer, and Michael S. Bernstein. LLM agents grounded in self-reports enable general-purpose simulation of individuals. *arXiv preprint arXiv:2411.10109*, 2024.

Ji-Lun Peng and Yun-Nung Chen. Rethinking role-playing evaluation: Anonymous benchmarking and a systematic study of personality effects. In *Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 205–218. Association for Computational Linguistics, 2026.

Letian Peng and Jingbo Shang. Codifying character logic in role-playing. In *Advances in Neural Information Processing Systems*, volume 38, pp. 16116–16147, 2025.

Letian Peng, Yupeng Hou, Kun Zhou, and Jingbo Shang. Codified finite-state machines for role-playing. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Letian Peng, Kun Zhou, Longfei Yun, Yupeng Hou, and Jingbo Shang. Deriving character logic from storyline as codified decision trees. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12480–12509. Association for Computational Linguistics, 2026b.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li. AgentSociety: Large-scale simulation of LLM-driven generative agents advances understanding of human behaviors and society. *iFuture*, 2026. doi: 10.26599/if.2026.9710004.

QwenLM. GitHub issue #525: reply by a Qwen maintainer on the latest date of the training data. https://github.com/QwenLM/Qwen3/issues/525, 2024. Reply of 11 June 2024, in the repository then named QwenLM/Qwen2; accessed 25 September 2026.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik R. Narasimhan, and Vishvak Murahari. PersonaGym: Evaluating persona agents and LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 6999–7022. Association for Computational Linguistics, 2025.

Nimrod Shabtay, Felipe Maia Polo, Sivan Doveh, Wei Lin, Muhammad Jehanzeb Mirza, Leshem Choshen, Mikhail Yurochkin, Yuekai Sun, Assaf Arbelle, Leonid Karlinsky, and Raja Giryes. LiveXiv – a multi-modal live benchmark based on arXiv papers content. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13153–13187. Association for Computational Linguistics, 2023.

Woojung Song, Nalim Kim, Sangjun Song, Chaewon Heo, Jongwon Lim, and Yohan Jo. ArcANE: Do role-playing language agents stay in character at the right time? *arXiv preprint arXiv:2606.05553*, 2026.

Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. CharacterEval: A Chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11836–11850. Association for Computational Linguistics, 2024.

Zeno Vendler. Verbs and times. *The Philosophical Review*, 66(2):143–160, 1957.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. FreshLLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13697–13720. Association for Computational Linguistics, 2024.

Lei Wang, Jianxun Lian, Yi Huang, Yanqi Dai, Haoxuan Li, Xu Chen, Xing Xie, and Ji-Rong Wen. CharacterBox: Evaluating the role-playing capabilities of LLMs in text-based virtual worlds. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6372–6391. Association for Computational Linguistics, 2025a.

Noah Wang, Z.Y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14743–14777. Association for Computational Linguistics, 2024a.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1840–1873. Association for Computational Linguistics, 2024b.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, Wei Wang, and Yanghua Xiao. CoSER: Coordinating LLM-based persona simulation of established roles. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 64822–64858. PMLR, 2025b.

Ziyi Wang, Yuxuan Lu, Wenbo Li, Amirali Amini, Bo Sun, Yakov Bart, Weimin Lyu, Jiri Gesi, Tian Wang, Jing Huang, Yu Su, Upol Ehsan, Malihe Alikhani, Toby Jia-Jun Li, Lydia Chilton, and Dakuo Wang. OPeRA: A dataset of observation, persona, rationale, and action for evaluating LLMs on human online shopping behavior simulation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 43942–43960. Association for Computational Linguistics, 2026.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. LiveBench: A challenging, contamination-limited LLM benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025.

Haotian Wu, Shufan Jiang, Chios Chen, Yiyang Feng, Hehai Lin, Heqing Zou, Yao Shu, and Chengwei Qin. FURINA: A fully customizable role-playing benchmark via scalable multi-agent collaboration pipeline. *arXiv preprint arXiv:2510.06800*, 2025a.

Xiaobao Wu, Liangming Pan, Yuxi Xie, Ruiwen Zhou, Shuai Zhao, Yubo Ma, Mingzhe Du, Rui Mao, Anh Tuan Luu, and William Yang Wang. AntiLeakBench: Preventing data contamination by automatically constructing benchmarks with updated real-world knowledge. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18403–18419. Association for Computational Linguistics, 2025b.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

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, Jian Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yaoning Yu, Ye Yu, Haojing Luo, and Haohan Wang. MiroBench: Benchmarking realism in agentic simulation of real-world discussions. *arXiv preprint arXiv:2606.14715*, 2026.

Terry Jingchen Zhang, Gopal Dev, Ning Wang, Max Obreiter, Punya Syon Pandey, Keenan Samway, Wenyuan Jiang, Yinya Huang, Bernhard Schölkopf, Mrinmaya Sachan, and Zhijing Jin. Test of time: Rethinking temporal signal of benchmark contamination. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 36538–36555. Association for Computational Linguistics, 2026.

Xinnong Zhang, Jiayu Lin, Xinyi Mou, Shiyue Yang, Xiawei Liu, Libo Sun, Hanjia Lyu, Yihang Yang, Weihong Qi, Yue Chen, Guanying Li, Ling Yan, Yao Hu, Siming Chen, Yu Wang, Xuanjing Huang, Jiebo Luo, Shiping Tang, Libo Wu, Baohua Zhou, et al. SocioVerse: A world model for social simulation powered by LLM agents and a pool of 10 million real-world users. *arXiv preprint arXiv:2504.10157*, 2025.

Qihao Zhao, Yangyu Huang, Tengchao Lv, Lei Cui, Qinzheng Sun, Shaoguang Mao, Xin Zhang, Ying Xin, Qiufeng Yin, Scarlett Li, and Furu Wei. MMLU-CF: A contamination-free multi-task language understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13371–13391. Association for Computational Linguistics, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023.

Jinfeng Zhou, Yongkang Huang, Bosi Wen, Guanqun Bi, Yuxuan Chen, Pei Ke, Zhuang Chen, Xiyao Xiao, Libiao Peng, Kuntian Tang, Rongsheng Zhang, Le Zhang, Tangjie Lv, Zhipeng Hu, Hongning Wang, and Minlie Huang. CharacterBench: Benchmarking character customization of large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24): 26101–26110, 2025a.

Lingfeng Zhou, Jialing Zhang, Jin Gao, Mohan Jiang, and Dequan Wang. PersonaEval: Are LLM evaluators human enough to judge role-play? In *Second Conference on Language Modeling*, 2025b.

Xinkai Zou, Yiming Huang, Zhuohang Wu, Jian Sha, Nan Huang, Longfei Yun, Jingbo Shang, and Letian Peng. Simulating organized group behavior: New framework, benchmark, and analysis. In *Third Conference on Language Modeling*, 2026.

A Benchmark Statistics

Table 3: Full build and evaluation split (716 representative and 284 fresh entities), both covering all 137 types. **Entities**: with at least one task. **People / fiction**: share of tasks from the *People* type and the four fictional-character types. **After**: targets starting on or after the oldest model's release.

	Entities	**Tasks**	**Tasks per entity**	**Ordered actions**	**People / fiction**	**Target dated to day / month / year / wider**	**Target year: median (5–95%)**	**After 2024-07-18**
Full build	18,360	438,792	13 / 563	969,088	2.7 / 2.4%	25 / 20 / 41 / 14%	1984 (1805–2023)	3.3%
Evaluation split	1,000	20,126	10 / 502	42,920	2.2 / 2.5%	26 / 22 / 39 / 12%	2001 (1841–2025)	11.5%

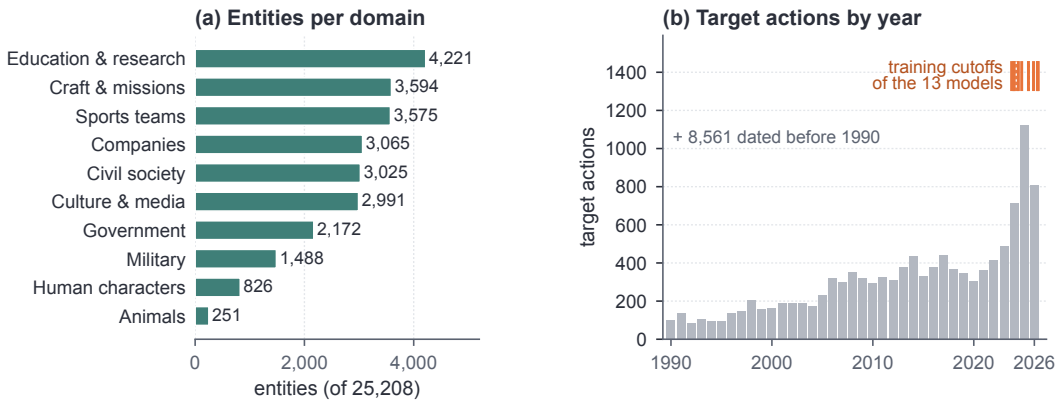


Figure 5: **Benchmark statistics.** (a) Entities per domain in the 25,208-entity benchmark; the 137 entity types are grouped into ten domains (Table 7). (b) Target actions of the 20,109 evaluated tasks by year; ticks mark the 13 models' training cutoffs.

B Evaluated Models

Table 4 lists the 13 evaluated models (§3.1): the checkpoint and decoding setting of each, its release date, and the training cutoff that the Live analysis uses. The vendors state nine of these cutoffs; the other four are tentative estimates (Appendix C).

Table 4: **Evaluated models.** Checkpoint, decoding setting, release date and the training cutoff used by the Live analysis (a cutoff stated as a month counts every action dated within that month as seen). †No vendor statement; tentative estimate (Appendix C). ‡The lowest setting GPT-6 Astra accepts; it spends 93 reasoning tokens per task on average, every other model none.

Model	Checkpoint	Setting	Release	Cutoff
OpenAI (API)				
GPT-4o	`gpt-4o-2024-11-20`	T=0	2024-11	2023-10
GPT-4.1-mini	`gpt-4.1-mini-2025-04-14`	T=0	2025-04	2024-06
GPT-5	`gpt-5-2025-08-07`	reasoning minimal	2025-08	2024-09
GPT-5.5	`gpt-5.5`	reasoning none, T=0	2026-04	2025-12
GPT-6 Astra	`gpt-6-astra`	reasoning low‡	2026-09	2026-04
GPT-6 Sol	`gpt-6-sol`	reasoning none, T=0	2026-09	2026-04
GPT-6 Luna	`gpt-6-luna`	reasoning none, T=0	2026-09	2026-05
DeepSeek (open weights; DeepSeek API)				
DeepSeek-V4-Pro	`deepseek-v4-pro`	thinking off, T=0	2026-08	2025-05
DeepSeek-V4.1-Flash	`deepseek-flash`	thinking off, T=0	2026-09	2025-05†
Qwen (open weights; local vLLM)				
Qwen2.5-72B	`Qwen/Qwen2.5-72B-Instruct`	T=0	2024-09	2023-12†
Qwen2.5-7B	`Qwen/Qwen2.5-7B-Instruct`	T=0	2024-09	2023-12†
Qwen3-8B	`Qwen/Qwen3-8B`	thinking off, T=0	2025-04	2024-04†
Meta (open weights; local vLLM)				
Llama-3.1-8B	`meta-llama/Llama-3.1-8B-Instruct`	T=0	2024-07	2023-12

C Tentative Training Cutoffs

The vendors of Qwen2.5-72B, Qwen2.5-7B, Qwen3-8B and DeepSeek-V4.1-Flash state no training cutoff. For each, Table 5 gives the value the Live analysis uses, its basis, and the other values we found and tested. Under every combination of the tested values, each model's drop after its cutoff stays above zero, with a 95% interval that excludes zero.

Table 5: **Tentative training cutoffs.** The vendors of four models state no training cutoff; the Live analysis uses the value in **Used**, the first day counted as after the cutoff. **Drop**: the model's score before − after the cutoff (as in Table 8) with the value used, and its range over the values tested. Over all 12 combinations of the values tested, every model's drop stays above zero with a 95% interval that excludes zero.

Model	Used	Basis of the value used	Other values tested	Drop	Range
Qwen2.5-72B	2024-01-01	"End of 2023": the only statement by Qwen staff (QwenLM, 2024), made in June 2024 for Qwen2, before Qwen2.5 was released	2024-05-01 and 2024-07-01, listed without a source by models.dev (2026) and OpenRouter (2026)	5.7	5.1 to 5.7
Qwen2.5-7B	2024-01-01	As Qwen2.5-72B	As Qwen2.5-72B	4.4	4.2 to 4.4
Qwen3-8B	2024-05-01	2024-04, measured on its sibling Qwen3-30B-A3B by HindsightBench (Jia, 2026)	2025-04-01, a quarter-end placeholder of OpenRouter (2026)	3.0	2.5 to 3.0
DeepSeek-V4.1-Flash	2025-06-01	DeepSeek-V4's official "approximately no later than May 2025" (DeepSeek-AI, 2026c), carried over	2026-01-01, a self-report in pre-release testing (linux.do, 2026)	6.3	4.6 to 6.3

D PROMPTS

Prediction. Every model receives the same single user message (no system prompt):

```
Below is a chronological list of things that happened
involving {entity} ({Wikidata class}).
1.  {date, or ``between A and B''} --- {action}
... (ten rows)
{question}
Answer with a single past-tense sentence describing one
action.  Do not explain.
```

Question. *What happened next with* {*acting party of the target*} *regarding* {*area*}*?* The area is a short generic noun phrase written by GPT-4.1-mini (temperature 0) from the target, with every name, title, number and date of the target withheld.

E JUDGE

For each prediction the judge answers five questions; the tier follows mechanically. (1) Does the prediction repeat one of the last two context actions instead of describing a new step? (2) Names aside, does it correctly describe the target, even if more vaguely? (3) Names aside, does it say what happened in the target, or one part of it? (4) Does it correctly give at least one specific of the target that the prompt does not give? (5) Does it fully cover the target, with nothing missing or incorrect? Tier 0 if (1) is yes or (2) is no; otherwise tier 1 if (3) is no; tier 4 if (5) is yes; tier 3 if (4) is yes; tier 2 otherwise. Questions (2) and (3) are read from a closed *kind* label the judge writes first (*same*, *too general*, *different*), not from its final yes/no.

F SCORE DISTRIBUTIONS

Table 6 gives, for each model, the share of tasks at each judge tier and the mean score with its 95% entity-bootstrap interval; these are the intervals of the All column of Table 2.

Table 6: **Judge tiers per model.** Share of the 20,109 tasks at each tier (%), and the mean score (25 × tier, 0–100) with its 95% entity-bootstrap interval (every entity counts once, as in Table 2). Rows are sorted by mean.

	Share of tasks (%) at tier						
Model	**0**	**1**	**2**	**3**	**4**	**Mean**	**95% CI**
GPT-6 Astra	24.9	3.5	23.2	18.3	30.1	55.1	[53.8, 56.3]
GPT-5.5	31.3	2.9	27.2	16.2	22.4	48.0	[46.7, 49.2]
GPT-6 Sol	30.7	6.6	29.3	14.8	18.6	45.6	[44.4, 46.8]
DeepSeek-V4-Pro	39.4	2.9	27.4	13.4	17.0	40.9	[39.7, 42.1]
GPT-6 Luna	38.6	3.7	29.3	12.6	15.8	40.6	[39.4, 41.8]
GPT-5	40.9	4.9	28.4	11.2	14.6	37.9	[36.8, 39.2]
GPT-4o	42.7	5.7	27.2	10.8	13.5	35.9	[34.7, 37.1]
DeepSeek-V4.1-Flash	44.6	5.2	26.5	10.2	13.5	34.4	[33.1, 35.7]
GPT-4.1-mini	44.0	6.8	29.5	8.5	11.3	33.3	[32.2, 34.4]
Qwen2.5-72B	46.7	11.2	27.0	6.1	8.9	28.0	[26.9, 29.1]
Qwen3-8B	53.0	6.6	29.0	4.4	7.0	26.1	[25.1, 27.2]
Llama-3.1-8B	55.6	2.8	27.8	6.2	7.6	25.3	[24.2, 26.4]
Qwen2.5-7B	57.1	8.1	24.5	3.7	6.6	23.9	[22.8, 24.9]

G Entity Domains

Table 7 lists the ten domains into which we group the 137 entity types for reporting (Table 2, Figure 2), with the entities and tasks of each in the evaluation split. *Human characters* combines people with literary, film, television and comics characters.

Table 7: **The ten entity domains.** The entity types each domain groups, with entities and tasks in the evaluation split.

Domain	Entity types	Entities	Tasks
Sports teams	American football teams, Baseball teams, Basketball teams, College sports teams, Cricket teams, Football clubs, Handball teams, Ice hockey teams, National football teams, National teams, Olympic delegations, Racing teams, Rugby teams, Sports clubs, Sports federations, Sports leagues, Volleyball teams, Wrestling promotions	142	2,291
Craft & missions	Aircraft, Battleships, Cargo ships, Cruisers, Destroyers, Expeditions, Frigates, Landing ships, Minesweepers, Museum ships, Ocean liners, Patrol boats, Sailing ships, Sailing warships, Satellites, Scientific projects, Ships, Software, Space missions, Space probes, Steamships, Submarines	173	3,438
Education & research	Boarding schools, Business schools, Colleges, Conservatories, Educational institutions, Engineering schools, Girls' schools, Hospitals, Law schools, Learned societies, Medical associations, Medical schools, Research institutes, Schools, Secondary schools, Seminaries, Student societies, Think tanks, Trade associations, Universities	139	1,706
Civil society	Advocacy groups, Charities, Churches, Clans, Community associations, Families, Monasteries, NGOs, Noble families, Political coalitions, Political parties, Religious orders, Religious organisations, Rights organisations, Social movements, Trade unions	115	1,708
Government	Cabinets, Commissions, Countries, Courts, Embassies, Emergency services, Government agencies, Governments, Intergovernmental bodies, Legislatures, Ministries, Municipal bodies, Tribal governments	95	1,249
Culture & media	Academic publishers, Bands, Broadcasters, Choirs, Dance troupes, Film distributors, Film studios, Game studios, Libraries, Museums, Music groups, Musical duos, Newspapers, Orchestras, Production companies, Publishers, Record labels, Theatre companies	130	3,210
Military	Air squadrons, Armed forces, Armed groups, Armies, Battalions, Criminal gangs, Military divisions, Regiments	53	2,194
Companies	Aerospace manufacturers, Airlines, Architecture firms, Banks, Breweries, Carmakers, Companies, Financial firms, Hospitality, Law firms, Manufacturers, Railway companies, Retail chains, Supermarkets, Technology companies, Transport companies	118	3,275
Animals	Animals	14	111
Human characters	Comics characters, Film characters, Literary characters, People, Television characters	21	927

H LIVE: PER-MODEL RESULTS

Table 8 gives the numbers behind Figure 3: each model's mean score on the tasks before and after its training cutoff, and the drop with its 95% entity-bootstrap interval. Tasks whose target straddles a model's cutoff are left out for that model (§2.3).

Table 8: Mean score (0–100) on tasks before and after each model's training cutoff, with Drop = Before − After in percentage points (computed before rounding) and its 95% entity-bootstrap interval; every interval excludes 0. †No vendor-stated cutoff; tentative estimate (Appendix C).

Model	Cutoff	Before	After	Drop (pp)	95% CI
GPT-4o	2023-10	38.0	28.5	9.6	[7.5, 11.7]
Llama-3.1-8B	2023-12	27.6	21.8	5.8	[3.8, 7.7]
Qwen2.5-72B	2023-12†	30.6	24.8	5.7	[4.0, 7.4]
Qwen2.5-7B	2023-12†	24.2	19.8	4.4	[2.5, 6.4]
Qwen3-8B	2024-04†	26.9	23.8	3.0	[1.1, 5.1]
GPT-4.1-mini	2024-06	35.1	26.9	8.2	[6.2, 10.4]
GPT-5	2024-09	39.4	31.0	8.3	[6.2, 10.8]
DeepSeek-V4-Pro	2025-05	42.1	35.2	6.9	[4.1, 10.1]
DeepSeek-V4.1-Flash	2025-05†	36.3	29.9	6.3	[3.4, 9.3]
GPT-5.5	2025-12	49.2	43.3	5.9	[2.3, 9.7]
GPT-6 Sol	2026-04	46.2	37.9	8.2	[3.3, 12.7]
GPT-6 Astra	2026-04	56.7	38.4	18.2	[12.3, 24.2]
GPT-6 Luna	2026-05	41.0	32.7	8.3	[1.7, 15.3]