A Supplement: Modeling Open-World Cognition as On-Demand Synthesis of Probabilistic Models

Appendix

A1 Related Work

The current work relates most closely to four lines of work, on model approximation and the Frame Problem, on LM guided model synthesis, hybrid models of language comprehension, and LM primitives in probabilistic programs.

A1.1 Model Approximation & the Frame Problem

Previous work on the Frame Problem has made significant progress in defining *resource rational* [34] objectives by which small, task-specific models can be constructed to approximate reasoning and planning with respect to larger models or priors [29, 27]. This work provides an important theoretical existence proof, demonstrating that it is possible to construct smaller tractable models that approximate larger (and intractable) ones, and that people empirically [27] show behaviors consistent with these approximations in reasoning and planning tasks. Both of these works have relatively little to say about *how* minds arrive at these smaller approximations. The current approach builds on this work by examining, at a Marr algorithmic level, how the mind might construct these models – by decomposing the process into a relevance-based synthesis procedure, and by showing that this can be instantiated concretely by exploiting learnable patterns acquired from joint program and language experience.

A1.2 Hybrid Models of Language Comprehension

Our concrete computational approach is more closely related to work in cognitive science that shows how language models can be used to synthesize probabilistic programs from language, by translating between natural language and a symbolic LoT [46, 52, 49, 51, 50]. This prior work considers cases where natural language explicitly spells out all relevant symbolic structure necessary for language interpretation. We build on these approaches by by extending model construction to areas where relevant knowledge must be recruited from large bodies of real-world background information, forcing us to confront the challenges of relevance-based retrieval that open-world reasoning poses.

A1.3 Language Models for Model Discovery

Our work connects to three related lines of work using code language models to synthesize structured models of the world or behavior. These lines of work differ in the goals of model synthesis, and the symbolic substrate of models they synthesize.

One thread focuses on using language models to synthesize explicit, symbolic computational cognitive models of human [40] or non-human animal behavior [9]. We differ from these works in our focus on synthesizing *probabilistic programs* as the key representational structure for representing ad-hoc models, which affords a particularly expressive model and automatic reasoning class with strong connections to earlier probabilistic modeling work in computational cognitive science. Our work is also somewhat different in its framing and goals. Both earlier works seek discover symbolic cognitive models to automate the proposal of scientific models for studying behavior. While our approach can be interpreted this way, the MSA architecture also represents an algorithmic hypothesis about *how humans minds actually reason*, framing flexible cognition itself as a process of ad-hoc model synthesis.

More broadly, our focus is on modeling how people reason about arbitrary, open-world situations differentiates – as a proof of concept towards more domain-general cognitive model synthesis over probabilistic models. This differentiates our work from other recent automated model synthesis methods in both cognitive science and AI that have focused on more domain-specific models, such as synthesizing models to explain social reasoning [53, 13]. Other recent AI work has focused on synthesizing world models that represent (often deterministic) transition functions for decision making and planning [44, 47, 41, 38]. This work could be productively combined with ours to synthesize probabilistic models that support planning and inference to explain an even wider class of ad-hoc reasoning.

Finally, a related and concurrent line of work in AI has begun to use language models to synthesize probabilistic models [18, 48], including probabilistic programs [33, 16]. These works are most similar to ours in their formalism, but differ significantly in their goals. The latter works especially focus on automating scientific modeling for statistical analysis from data. We focus on an expressive probabilistic programming language class designed for cognitive modeling, and evaluate our approach with respect to empirical evidence of human reasoning. However, as with other work on automated modeling, there are rich synergies between these approaches – such as extending the MSA approach to capture human scientific discovery, or collaborative scientific discovery between AI and human "thought partners" that includes jointly modeling a human scientist along with models of the world [11].

A2 General Discussion

A2.1 Reasons For Model Fit

Why might human judgments better align with MSAs than with LM baselines in these cases? One possibility is that the difference is due to the way both model classes handle coherence. The mental models generated by MSAs are coherent by design, while LMs internal representations do not have similar coherence constraints. If people's judgments over multiple variables tend to be more internally coherent, this could drive the fit to MSAs over LMs. Another possibility is that MSAs use of explicit causal and probabilistic representations might force the model to place more weight on deeper structural properties, rather than superficial features of the language used to describe tasks. If people's judgments are tracking these deeper causal properties of the stimuli, this could explain the better match to MSAs. Such an explanation would fit with similar findings that point to a lack of robustness in these models in response to surface-level features [36, 43, 37]. Determining which of these or other explanations is most plausible, and if this general trend continues to hold in more varied domains, is a priority for future research.

A2.2 Handling Surprising Evidence

In our data, people appeared to be close to rational in their integration of evidence with background beliefs, as measured by fit to our MSA. This included integrating unexpected observations (e.g., a surprising win by a suspected slow runner against a suspected fast runner) in a measured fashion. In cases where LMs differ most from people, a tentative analysis suggests that one of the key challenges faced by LMs was an over-sensitivity to these surprising observations. For example, from qualitative inspection, we noticed instances where the LM baselines tended to believe that a fast runner's single loss to an otherwise slower runner was often enough to neutralize or reverse the model's assessment of their relative speeds, even when the weight of the rest of the evidence suggested otherwise. The tendency of our MSA not to over-index in these cases may be due to the construction of an explicit model, with priors and a causal structure that grounds the integration of competing observations. Further work should explore this theme of holistic integration more thoroughly, including in cases where information is revealed piecemeal over time (as it often is in naturalistic reasoning tasks), rather than all at once (as in our experiments here).

A2.3 Open-World Reasoning

Data from Exp. 3 demonstrated the largest differences between model classes in fit to human data. This experiment focused on generalization in the open-world setting, conditioning on participant-sourced commentaries introducing novel considerations. Performance on this experiment represents a particularly interesting kind of generalization – to observations that require introducing new variables and dependencies into the underlying causal structure, thereby expanding the expressivity of the model (relative to what would have been synthesized in the absence of the commentary; the models synthesized in Exp. 1 and Exp. 2). As noted earlier, reasoning in this open-world setting represents a strong challenge for classical Bayesian models of cognition, which cannot handle novel variables. Despite this, our MSA strongly outperforms LMs in modeling human judgments for these stimuli, suggesting a continued benefit from being able to rely on the kinds of representations that figure in probabilistic models. In particular, MSAs' ability to recombine symbolic representations of the relevant causal structures may have supported a greater degree of generalization to highly novel circumstances. A priority for future work is explore where this ability breaks with LM-powered model synthesis to explore whether other kinds of MSAs might better fit human cognitive abilities in turn.

A2.4 Distributions of Human Judgments Reveal Structure That All Models Fail To Capture

One of the advantages of collecting and analyzing distributional data is that we can analyze human and model judgments in more fine-grained ways than conventional measures like R^2 allow. A cursory analysis of this data reveals interesting differences between people and both model classes, highlighting the amount of structure in human judgments still to be explained. Compared to human participants, for example, LMs appear to be more streaky – clustering their judgments around particular outcomes – and respond too strongly to surprising observations – yielding judgments that

are at times wholly in the opposite direction of people's. MSAs are more often directionally correct (as evidenced by greater R^2), but tend to produce judgments that are visibly smoother and more uncertain than people's (see Figure 15 in Supplement). In short, human judgments appear to have strong opinions (visible streaks, like in the LMs), but place those peaks more consistently over modes in the Bayesian posterior (as evidenced by our MSA's superior R^2 and WD measures).

A pressing question then is whether some other model class could better delivers the patterns seen in the human data. This might be some deeper hybridization of neural and symbolic methods – one that reproduces the sharply peaked opinions of LMs, but places those peaks more consistently in the right places – or an MSA with stronger sampling methods that focus samples more directly over modes. Modeling such fine-grained distributional features of human judgments is a target for future work.

A2.5 Experimental Limitations

One limitation of the current work is that human data were relatively noisy – both split half human-human correlations and model-human correlations showed wide confidence intervals. We can also explore ways to make human vairance more model-able – by matching particular mental models (in MSAs) or response patterns (in LMs) to particular participants – to capture individual participant's unique conception of the situtation, for example.

Variance in samples from our MSA was also often too low. Judgments in Experiments 1 and 2, for example, were highly correlated for our MSA, but not nearly so correlated for people. Similarly, people's judgments in certain conditions, such as in the canoe domain, were often higher variance than those of our MSA. This suggests a lack of diversity in the models synthesized by our MSA. Follow-on work should explore how to increase the diversity of synthesized models, by increasing the number of models, for example, or by more targeted methods, such as conditioning model generation (and LM responses) on participants' self-reports about what they are thinking about.

Another near-term target for follow-up work is exploring stronger baselines and more thorough model ablations. Anecdotally, we found that a staged model synthesis procedure worked best, but this should be explored systematically and compared to other model synthesis strategies. Similarly, MSA performance should be compared to state-of-the-art reasoning models, as well as the cognitive models derived from them [6]. Leading reasoning models in particular are likely to perform better at these tasks, but also likely to synthesize better probabilistic models if used internally to our MSA. It will be important to see how those two effects wash out when both are compared for human-likeness.

A2.6 Limitations in Model Synthesis

Much like the LMs, our implementation of an MSA also faced important limitations in its ability to generalize. Model generations were often overly influenced by the example models given in our prompt, with a consequent lack in model diversity. For example, while our MSA was often able to reconfigure the primitives in the prompted models into models for the novel sport, it struggled to invent new primitives when these were called for. In Exp. 3 our MSA struggled to make sense of temporal information frequently given in commentaries (e.g., "Kai was fast until he rolled his ankle in match 4") until we included an example of the relevant abstraction, a temporal ordering of events, in the prompted models. Once armed with this abstraction, the MSA could model the influence of events before, after, or during, but it struggled to build these abstractions on its own. Some of these issues of prompt sensitivity might be ameliorated by using larger LMs or models specially fine-tuned for the task of model synthesis, which might learn to more systematically explore the space of possible models.

A2.7 Looking Ahead

Addressing the problem of open-world cognition that will require exploring a broader space of possible MSAs. This might include synthesis using other modeling languages that support long-horizon planning [54], multi-agent reasoning [10], or distributional primitives learned from experience [32, 15, 25]. Future work should also explore other model synthesis strategies, such as those that refine initial models with external feedback [47, 44] or that consider multiple models at once [35]. Finally, future MSAs should *learn* from model construction over time, by components of the synthesis architecture based on previous successes or failures, and by augmenting the modeling language with successful concepts [17, 24].

Future MSAs can be evaluated both based on ground truth accuracy – whether the models they synthesize are any good – and match to various measures of human behavior. We can ask, for example, whether certain modeling languages better capture the generalizations that people endorse, or which synthesis strategies fit the dynamics of human thought processes, as measured by reaction times or systematic shifts in people's judgments.

The current era of highly general AI systems means that a deep understanding of how human open-world cognition works may now be within reach. We don't yet have a settled view of how people are able to reason in locally coherent and globally relevant ways about the large and ever-expanding space of things people think about, but the way to investigate this is becoming clear. By scaling MSAs, as well as their pure LM alternatives, and systematically comparing them to human data, we can now begin to meaningfully adjudicate between models of human general cognition. Cognitive science has shed tremendous light on how parts of the mind work. It can now begin to study how those parts fit together.

A3 Model Synthesis Architectures: Additional Implementational Details

Experiment and model implementation details reference the repository at: `https://anonymous.4open.science/r/msa-cogsci-2025-data-CFB6`.

As described in the main text, we sequentially construct $M_{\text{ad-hoc}}$ in a staged process that interleaves generation and evaluation steps. The base LM used in all experiments is the HuggingFace `meta-llama/Meta-Llama-3.1-70B-Instruct-Turbo` release. We query the model using the Together API. Here we provide additional parameters and prompting details for each of these stages.

In our experiments, as we described in the main text, we model each *simulated human participant* as ultimately synthesizing a single model $M_{\text{ad-hoc}}$ conditioned on an input natural language scenario. The following describes the parameterization used for each single simulated human participant.

Each stage of generation involves a frame prompt for that stage, into which we inject a shuffled set of background examples demonstrating each stage of this pipeline for a set of held-out example scenarios (none of which appear verbatim in our main experiments.) Specifically, we use a *held-out* prompting scheme for selecting these examples, where for a scenario from any given sporting domain (eg. *tug-of-war*) we automatically select background examples only constructed for the *other* sports – in this case, *canoe-racing* and *biathlon*, along with two other example scenarios, *diving* and *exam*, that we use as examples for all scenarios.)

Below, we describe where in the repository one can find the *frame prompts* for each stage, which include a `<SHUFFLED EXAMPLES>` token indicating where these shuffled example generations appear. The full set of shuffled examples themselves can be found at the `example-scenarios` directory at our data repository, which includes:

- Base {`tug-of-war, canoe-racing, biathlon, diving, exam`} examples used for **Exp. 1** and **Exp. 2**.
- Base {`tug-of-war, canoe-racing, biathlon, diving`} examples for **Exp. 3**. This experiment was run later and we constructed extended examples demonstrating models with free-form additional natural language observations. We also omit the exam example domain from these experiments. However, future work will explore the effect of these examples on generation and seek to construct a more general set of examples (or fine-tune models so that example-based prompting is not necessary; we use it here as we build on a generic base model.)

Note that these shuffled example text files contain a concatenated set of *all* of the generation stages (eg. each example file contains an example input scenario, parse, background information in natural language, dependency graph, and full probabilistic program.

All frame prompts for each generation stage appear under the `msa-frame-prompts` directory. Generating of the parsing and background-knowledge/dependency graph used a single system prompt

which is included in the same directory. No system prompt was used for the ad-hoc probabilistic program model generation stage.

A3.1 Parsing

In our experiments, we forward sample only $k_{parse} = 1$ parse at `temp=0.2`. Throughout, we use lower temperatures for generation stages that require greater syntactic control (like code generation) and higher temperatures for tasks that involve generating natural language (like retrieving and generating informal relevant variables.) We also implement an LLM-based evaluation function Φ_{parse} which scores parses, but as we only take $k_{parse} = 1$ sample per participant this is of limited utility (we find empirically that parse variability is less important for downstream model quality than diversity in informal knowledge generation, but k_{parse} could be increased for more ambiguous and freeform language in future experiments.

The full frame prompt for the parsing stage can be found at `generate-parsing` in the frame prompts directory and the evaluation prompt can be found at `score-parsing`.The frame prompt for this stage was injected with shuffled and concatenated examples starting from the input scenario up to the example parses (delimited by `<START_LANGUAGE_TO_WEBPPL_CODE>`).

Here we show a few example parses for canoe-racing and biathlon scenarios in **Exp. 1** and **Exp. 2** (as only the background information changed between these experiments, the outcome evidence and questions shown were matched for scenarios in Exp. 1 and Exp. 2). We omit a tug-of-war example as the latent variables as it uses the same outcome and latent variable format as canoe-racing. Parses are excerpted from the full scenario, but show examples of a sentence in natural language parsed into a corresponding line of code. Note that the parse code invariably includes calls to placeholder functions that have not yet been generated and must be generated in the final model.

Example parse for Exp. 1, canoe-racing

In the first race, Fey and Ollie lost to Lane and Jamie.

```
condition(lost({team1: ['fey', 'ollie'], team2: ['lane', 'jamie'], race: 1}))
```

Query 1: Out of 100 random athletes, where do you think Fey ranks in terms of intrinsic strength?

```
intrinsic_strength_rank({athlete: 'fey', out_of_n_athletes: 100})
```

On a percentage scale from 0 to 100%, how much effort do you think Fey put into the second race?

```
effort_level_in_race({athlete: 'fey', race: 2})
```

In a new race later this same day between Fey and Ollie (Team 1) and Harper and Gale (Team 2), who would win and by how much?

```
who_would_win_by_how_much({team1: ['fey', 'ollie'], team2: ['harper',
    'gale'], race: 4})i
```

Excerpted parse examples for Exp. 1, biathlon

In the first round, Robin and Ollie beat Lane and Ness.

```
condition(beat({team1: ['robin', 'ollie'], team2: ['lane', 'ness'], round:
    1}))
```

Out of 100 random athletes, where do you think Robin ranks in terms of intrinsic strength?

```
intrinsic_strength_rank({athlete: 'robin', out_of_n_athletes: 100})
```

On a percentage scale from 0 to 100%, how accurate do you think Robin was at shooting in the second round?

```
shooting_accuracy_in_round({athlete: 'robin', round: 2})
```

In a new round later this same day between Robin and Ollie (Team 1) and Lane and Taylor (Team 2), who would win and by how much?

```
who_would_win_by_how_much({team1: ['robin', 'ollie'], team2: ['lane',
    'taylor'], round: 4})
```

Here we show a few excerpted example parses for **Exp. 3**, specifically showing parses of the free-form participant-provided observations. As there is more variability across these parses, we show several instances of parses sample for different simulated participants to demonstrate variability.

Example parses for Exp. 3, participant-generated details

Taylor is brand new to the sport of canoe racing, and this is only his 2nd time competing.

```
Sampled parse 1: condition(is_brand_new_to_canoe_racing({athlete: 'taylor'})
    && is_second_time_competing({athlete: 'taylor'}))

Sampled parse 2:  condition(is_brand_new_to_canoe_racing({athlete:
    'taylor'}) && is_only_second_time_competing({athlete: 'taylor'}))
```

Kay didn't get enough sleep last night and can barely stay awake during the race.

```
Sampled parse 1: condition(!got_enough_sleep_last_night({athlete: 'kay'}) &&
    barely_staying_awake_during_race({athlete: 'kay'}))

Sampled parse 2: condition(didnt_get_enough_sleep_last_night({athlete:
    'kay'}))
```

In the first match of tug-of-war Kay, while managing to pull off the win from Avery, pulled a muscle in their shoulder that limited their pulling output going forward.

```
Sampled parse 1: condition(pulled_muscle_in_shoulder_in_match({athlete:
    'kay', match: 1}) && beat({team1: ['kay'], team2: ['avery'], match: 1}))

Sampled parse 2: condition(pulled_muscle_in_match({athlete: 'kay', match:
    1}) && beat({team1: ['kay'], team2: ['avery'], match: 1}))
```

A3.2 Retrieving informal relevant background knowledge and proposing conceptual dependency graph

We jointly sample the informal relevant background information K and corresponding dependency graphs G. In our experiments, for each simulated participant we sample $k_{informal} = 8$ informal specifications and their corresponding dependency graphs at $temp = 0.5$. We then implement an LLM-based evaluation function $\Phi_{informal}$ which jointly scores the generated K and G, from which we select the top scoring $K*, G*$.

The full frame prompt for this stage can be found at `generate-informal-background-knowledge-and-dependency-graph` and the evaluation prompt can be found at `score-informal-background-and-dependency-graph` in the frame prompts directory. As each generation stage is conditioned on all previous generation steps, note that the injected shuffled and concatenated examples now draw from the input scenario up to the example depdendency graph (delimited by `<START_SCRATCHPAD>`).

Here we show example retrieved informal background knowledge and the corresponding dependency graph for sample scenarios in **Exp. 2** (which required retrieving additional information to make up for the underspecified background) and **Exp. 3**.

Example informal background knowledge and dependency graph for Exp. 2, biathlon

In this event, teams of players are competing in rounds of a biathalon, a winter sport that combines cross-country skiing races and rifle shooting. In each round, the team that wins depends on the average speed with which the athletes are able to ski, based on their intrinsic strength, as well as each team member's shooting accuracy in that particular round.
Intrinsic strength is an underlying attribute of a given athlete that varies somewhat widely from athlete to athlete. An athlete's intrinsic strength ranking out of n other athletes is the number of other athletes we might expect them to be stronger than out of N total random athletes.
Athletes also vary in their shooting accuracy in any given round. Shooting accuracy is a continuous parameter, measured as a percentage from 0 to 100%. Athletes can have poor, average, or excellent shooting accuracy in a given round. Their shooting accuracy in a round is somewhat dependent on their intrinsic strength, as stronger athletes are probably more likely to have better shooting accuracy.
A team's skiing speed in a round is described in the background as the average effective skiing speed with which the athletes are able to ski on that team. The effective skiing speed of an athlete in a round is determined at a base level by their intrinsic strength, since stronger athletes ski faster.
A team's overall score in a round is the sum of their average skiing speed and their average shooting accuracy in that particular round. The skiing speed and shooting accuracy are weighted equally.
The team that gets the highest score in any given round wins. A team beats another if their team's overall score is higher in that round. A team loses to another team if they did not beat that team in a given round.
The amount that one team wins over another team is given by the difference in their overall scores.

```
- intrinsic_strength
- intrinsic_strength_rank
  - depends on: intrinsic_strength

- shooting_accuracy_in_round
  - depends on: intrinsic_strength

- effective_skiing_speed_in_round
  - depends on: intrinsic_strength

- team_skiing_speed_in_round
  - depends on: effective_skiing_speed_in_round

- team_shooting_accuracy_in_round
  - depends on: shooting_accuracy_in_round

- team_overall_score_in_round
  - depends on: team_skiing_speed_in_round, team_shooting_accuracy_in_round

- beat
  - depends on: team_overall_score_in_round
- lost
  - depends on: beat
- who_would_win_by_how_much
  - depends on: lost
```

Example informal background knowledge and dependency graph for Exp. 3, tug of war

Participant-generated detail: In the first match of tug-of-war Kay, while managing to pull off the win from Avery, pulled a muscle in their shoulder that limited their pulling output going forward.
First, let's reason about the role of strength in this scenario. Intrinsic strength is an underlying attribute of a given athlete that varies somewhat widely from athlete to athlete.
An athlete's intrinsic strength ranking out of n other athletes is the number of other athletes we might expect them to be stronger than out of N total random athletes.
Athletes also vary in the effort that they put into any given match. Athletes can put in either moderate amount of effort, little effort, or extra high amounts of effort. Which of these they are more likely to do probably depends on their underlying strength, as stronger athletes are probably more likely to put in extra high effort, and weaker athletes probably tend to be more likely to put in lower amounts of effort.
An athlete who 'tries hard' in a match puts in a fair amount of effort.
Whether or not an athlete pulls a muscle in their shoulder in a specific match occurs at a rare frequency for any given athlete and match. This is an event that affects future matches after the match in which someone was injured. We will need to think about whether an athlete has pulled a muscle in ANY previous matches to understand its effects on the current match.
An athlete's effective pulling strength in a given match is determined at a base level by their intrinsic strength, but is (1) reduced if they pulled a muscle in their shoulder in any PREVIOUS match, which will make them pull less hard; and (2) increased by their effort level, which is effectively a percentage multiplier on their pulling strength in this match.
A team's pulling strength in a match is described in the background as the AVERAGE effective pulling strength with which the athletes are able to pull on that team.
A tug-of-war team beats another if their team's pulling strength is greater in that match, assuming a fixed match length.
A tug-of-war team loses to another team if they did not beat that team in a given match.
To calculate who would win and by how much, we will calculate the likelihood that a team would win over another.

```
- intrinsic_strength
- intrinsic_strength_rank
  - depends on: intrinsic_strength
- effort_level_in_match
  - depends on: intrinsic_strength
- pulled_muscle_in_shoulder_in_match
- pulled_muscle_in_shoulder_in_any_previous_match
  - depends on: pulled_muscle_in_shoulder_in_match
- effective_athlete_pulling_strength_in_match
  - depends on: intrinsic_strength,
    pulled_muscle_in_shoulder_in_any_previous_match, effort_level_in_match
- team_pulling_strength_in_match
  - depends on: effective_athlete_pulling_strength_in_match
- beat
  - depends on: team_pulling_strength_in_match
- lost
  - depends on: beat
- who_would_win_by_how_much
  - depends on: lost
```

A3.3 Synthesizing the formal symbolic model

Finally, conditioned on the best scoring parse, informal knowledge and dependency graph, for each simulated participant we sample a single full probabilistic program $k_{program} = 1$ at $temp = 0.2$ from the LLM. As described in the main text, we implement Φ_{formal} as simply whether or not the probabilistic program compiles and produces inferences over the queries.

The full frame prompt for this stage can be found at `generate-model` in the frame prompts directory. As each generation stage is conditioned on all previous generation steps, note that the injected shuffled and concatenated examples now include the full text of the examples, from the scenarios and including all prior example generation stages up to the example models.

As models are verbose, below we reproduce an *excerpted* model from **Exp. 3** showing the parse that involves functions for the new participant detail, and with excerpted components for the other named functions.

Example informal background knowledge and dependency graph for Exp. 3, biathlon

Participant-generated detail: In the first match of tug-of-war Kay, while managing to pull off the win from Avery, pulled a muscle in their shoulder that limited their pulling output going forward.

```
var model = function() {
// BACKGROUND KNOWLEDGE

    // Intrinsic strength is an underlying attribute of a given athlete that
    varies somewhat widely from athlete to athlete.
    var intrinsic_strength = mem(function ({athlete}) {...})

    // An athlete's intrinsic strength ranking out of n other athletes is
    the number of other athletes we might expect them to be stronger than
    out of N total random athletes.
    var intrinsic_strength_rank = function({athlete, out_of_n_athletes})
    {...}

    // Athletes also vary in the effort that they put into any given match.
    Athletes can put in either moderate amount of effort, little effort, or
    extra high amounts of effort. Which of these they are more likely to do
    probably depends on their underlying strength, as stronger athletes are
    probably more likely to put in extra high effort, and weaker athletes
    probably tend to be more likely to put in lower amounts of effort.
    var effort_level_in_match = mem(function({athlete, match}) {...})

    // An athlete who 'tries hard' in a match puts in a fair amount of
    effort.
    var tried_hard_in_match = mem(function({athlete, match}) {...})

    // Whether or not an athlete pulls a muscle in their shoulder in a
    specific match occurs at a rare frequency for any given athlete and
    match. This is an event that affects future matches after the match in
    which someone was injured. We will need to think about whether an
    athlete has pulled a muscle in ANY previous matches to understand its
    effects on the current match.
    var pulled_muscle_in_shoulder_in_match = mem(function({athlete, match}) {
      var likelihood_of_pulling_muscle_in_match = 0.05;
      return flip(likelihood_of_pulling_muscle_in_match);
    })

    // An athlete's effective pulling strength in a given match is
    determined at a base level by their intrinsic strength, but is (1)
    reduced if they pulled a muscle in their shoulder in any PREVIOUS match,
    which will make them pull less hard; and (2) increased by their effort
    level, which is effectively a percentage multiplier on their pulling
    strength in this match.
    var effective_athlete_pulling_strength_in_match = mem(function({athlete,
    match}) {
      // Assume that base pulling strength is just their current strength.
      var base_pulling_strength_in_match = intrinsic_strength({athlete :
    athlete})

      //  Reduced if they pulled a muscle in their shoulder in any PREVIOUS
    match. Use the helper function to check if they pulled a muscle after
    any previous match.
      var pulling_strength_adjusted_for_pulled_muscle =
    any_previous_time_inclusive(
        function(prev_match) {
          return pulled_muscle_in_shoulder_in_match({athlete: athlete,
    match: prev_match})
        }, match) ? base_pulling_strength_in_match * 0.7 :
    base_pulling_strength_in_match;

      // Increased by effort level in this match.
      var pulling_strength_adjusted_for_effort_level =
    (effort_level_in_match({athlete: athlete, match: match}) / 100) *
    pulling_strength_adjusted_for_pulled_muscle;

      return pulling_strength_adjusted_for_effort_level;
```

A3.4 Model-Based Bayesian Inferences

In our experiment, all inferences are derived using the WebPPL built in *rejection sampling* inference engine. Inference budgets are specified in the main text: we report posteriors from $b_{samples} = 1000$ samples per simulated participant for **Exp. 1** and **Exp. 2**, and $b_{samples} = 500$ for **Exp. 3** (as in general rejection sampling is much slower on these, where observations specify a rare a priori observation.

A4 Natural Language Reasoning Experiments: Additional Experimental Details

A4.1 Model Olympics Vignettes

This supplemental section provides additional details on the stimuli generation and selection process for the Model Olympics domain vignettes used throughout the experiments.

As described in the main text, we construct a set of procedurally generated vignettes for experiments **Exp. 1, Exp. 2, Exp. 3**, where each vignette consists of a: linguistic *background* on the particular sport of interest (which could be *tug-of-war*, *canoe racing*, or *biathlon*); a set of *evidence* sentences describing match outcomes (plus, in the **Exp. 3** case, one additional participant-generated observation); and 8 *questions*.

At the data repository section, the `model-olympics-human-experiment` directory contains:

- Base **detailed backgrounds** for the {`tug-of-war, canoe-racing, biathlon`} sports used for vignettes in **Exp. 1**.
- Base **underspecified backgrounds** for the {`tug-of-war, canoe-racing, biathlon`} sports used for vignettes in **Exp. 2**.
- Base **underspecified backgrounds (no reference to any participant-generated variables)** for the {`tug-of-war, canoe-racing`} sports used for vignettes in **Exp. 3**. Note that the Exp. 3 vignettes were constructed shown to models were constructed using underspecified backgrounds (as in Exp. 2); we provide these again for comparison.

Using the base backgrounds, we procedurally generated vignettes for each sport using a set of **16 base vignette templates**, comprised of **12** templates derived from the patterns of evidence used (originally, in the tug-of-war domain only) in [22], and **4** additional templates specifically designed to present noisy and anomalous evidence that would evaluate whether participants and models judged these outcomes based on the "Bayesian explaining away" of anomalous outcomes relative to accumulative contrary evidence, based on multiple conjunctive latent causal variables. The templates describe the relations between athletes in a tournament; we instantiate the templates into concrete templates for each sport using sport-specific latent variables, and with randomly sampled athlete names from a set of gender-neutral names (to avoid priors about athlete strength). We then randomly subsampled amongst these procedurally templates to select the vignettes reported in our experiments. In total, as described in the main experimental text the final stimuli for each experiment reported in the paper comprised:

- **Exp. 1**: 6 randomly sampled vignettes (from the full set of 16 possible vignette templates) for each sport, for a total of **18** vignettes. Note that these 6 vignette templates were independently sampled for each sport and therefore may not have had the same evidence patterns per sport.
- **Exp. 2:** matched vignette templates to Exp. 1, for a total of **18** vignettes, but with underspecified backgrounds and re-generated athlete names.
- **Exp. 3:** 5 tug-of-war and 4 canoe-racing vignettes, which were base vignettes extended with participant-generated details. As we describe throughout, these base vignettes were similar in form but slightly different (in their background details and phrasing of the inference questions) from those used in Exp. 1 and Exp. 2, as this was a preliminary experiment piloted before Exp. 1 and Exp. 2.

As we describe in the human experimental details below, human participants in our study actually viewed slightly more vignettes than were ultimately reported in our paper here or compared to model results – we removed one vignette (from all three reports) which contained an error in the questions that asked about an athlete who was actually not part of a particular match; and we removed one additional sports domain, a synchronized*diving* domain, due to apparent confusion about the sport itself and high amounts of variance in participant answers. We currently withhold the full exact set of stimuli from **Exp. 1** and **Exp. 2**, and the stimuli used in **Exp. 3**, to avoid their appearance in LLM training datasets while we prepare an extended version of this work. The full dataset will be released upon publication, and future work will seek to generate a more dynamic version of this dataset for evaluation. However, here we show an example vignette from the *tug-of-war* domain, demonstrating the difference between the detailed (Exp. 1) and underspecified (Exp. 2) backgrounds.

Example tug-of-war vignette, showing Exp. 1 vs. Exp. 2 backgrounds

Exp. 1: Detailed background In this event, the athletes are competing in tug-of-war tournaments. Each tournament consists of a series of matches. In each match, athletes compete as part of a team.
An athlete's intrinsic strength remains constant throughout a tournament. An athlete neither gets stronger nor weaker between matches. You can assume that all matches take place on the same day.
Athletes also vary in the effort that they put into any given match. Most of the time, people pull with a moderately high amount of effort. Sometimes, an athlete won't put in much effort and will pull with only a fraction of their strength. Other times, they may put in a lot of effort and pull extra hard, beyond what their intrinsic strength would suggest.
How hard a team pulls overall in any given match is determined by the total amount that all of the athletes on the team pull in that match. How hard each athlete pulls in a given match is determined by their intrinsic strength, modified by how much effort they put in (a lower fraction of their intrinsic strength if they don't put in much effort, or even more than their strength if they put in more effort).
The team that pulls the hardest in a given match wins.
Athletes compete either individually or as a team.
All matches take place on the same day.

Exp. 2: Underspecified background In this event, the athletes are competing in matches of tug-of-war.
In each round, the team that wins the round depends on how hard the athletes collectively pull, based on their intrinsic strength modulated by other factors including how much effort they put in to that round.
Athletes compete either individually or as a team.
All matches take place on the same day.

CONDITIONS
In the first match, Peyton and Avery lost to Blake and Casey.
In the second match, Peyton and Blake lost to Avery and Casey.
In the third match, Peyton and Casey lost to Avery and Blake.

QUERIES
Query 1: Out of 100 random athletes, where do you think Peyton ranks in terms of intrinsic strength?
Query 2: Out of 100 random athletes, where do you think Avery ranks in terms of intrinsic strength?
Query 3: Out of 100 random athletes, where do you think Blake ranks in terms of intrinsic strength?
Query 4: On a percentage scale from 0 to 100%, how much effort do you think Peyton put into the second match?
Query 5: On a percentage scale from 0 to 100%, how much effort do you think Avery put into the second match?
Query 6: On a percentage scale from 0 to 100%, how much effort do you think Blake put into the second match?
Query 7: In a new match later this same day between Peyton and Avery (Team 1) and Blake and Gale (Team 2), who would win and by how much?
Query 8: In a new match later this same day between Peyton and Blake (Team 1) and Avery and Gale (Team 2), who would win and by how much?

A4.2 LM-only experimental details

The repository includes the frame prompting format used to elicit judgments for both the *LM-direct* and *LM-CoT* baselines, in the `lm-only-baseline-prompts` directory.

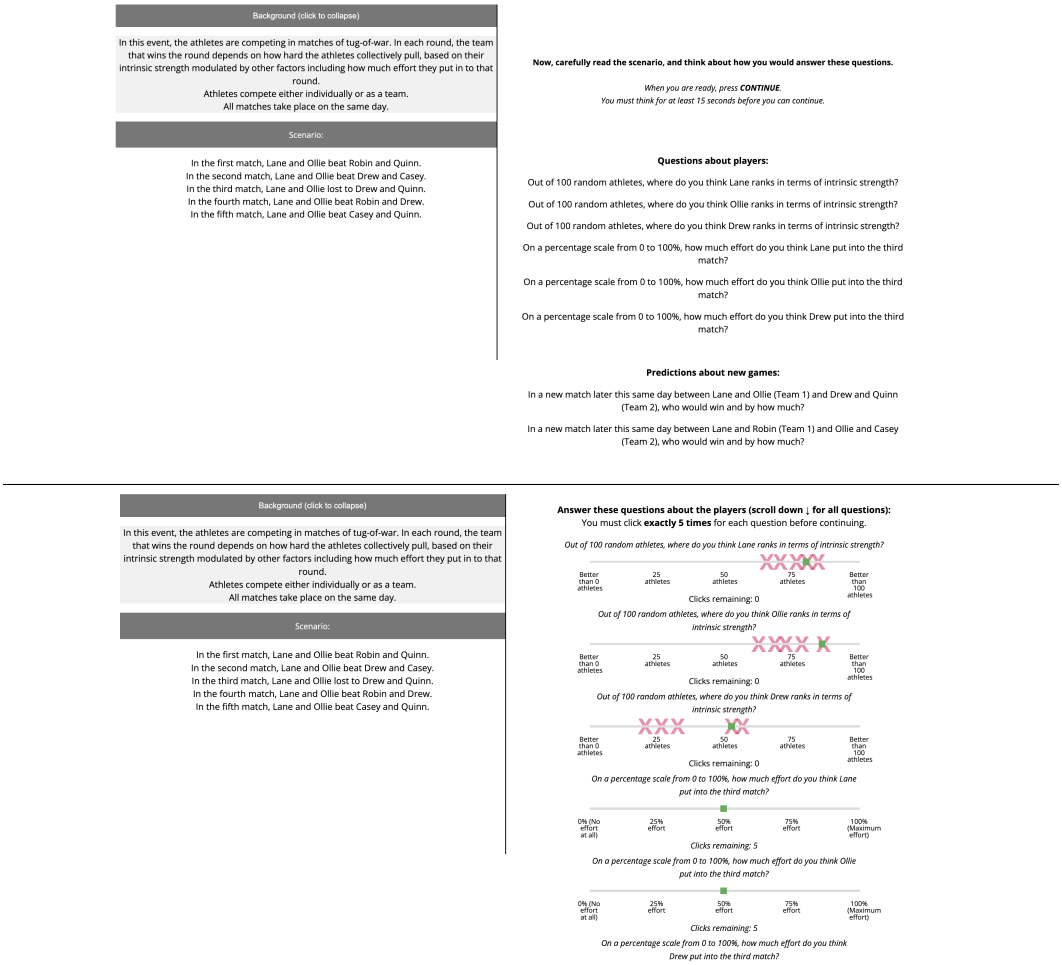

Figure 7: Example interfaces showing the human experimental setup – shown is a sample trial from **Exp. 2**, the underspecified background experiment. Participants first read the background information, scenario and questions (top). They then indicate their judgments via multiple clicks per question (bottom)

Each prompt contained the full *experiment instructions* shown to humans for each experiment (though note that the videos showing how to use the multi-click judgment interface were described in text, as prompts were text only); and then the full vignette, with additional instructions for how to answer each query.

A4.3 Exp. 1: human judgment experimental details

Pre-trial instructions were shown to all participants containing an example tutorial on how to use the multi-click slider interface to indicate the distribution of their judgments, including GIFs showing how they could indicate high certainty about a specific posterior mode (eg. most clicks around one end of the slider); split certainty about multiple posterior modes; and relative uncertainty about a continuous range.

As described in the main text, participants judged a randomly constructed batch of two vignettes from each of the three sport. Trials were grouped by sport (all participants first saw vignettes about tug-of-war, then canoe-races, then biathlon. Participants in 3 of the 4 batch conditions (57 of 76 participants) also saw an additional 2 vignettes from an additional synchronized diving domain; this domain was moved after participants appeared generally confused about the domain and showed extremely low inter-participant correlation in answers.

Trials were grouped into sections by each sport. Prior to reading the vignettes for a particular sport, in Exp. 1 only, participants were additionally presented with a full background description of the sport (with the same details as in the *detailed background* D(, along with an example tournament showing the kinds of outcome patterns that could appear in the later vignettes). Participants were required to spend 15 seconds reading this background description. The full text of these background descriptions can be found in the `model-olympics-human-experiment` directory of our repository. Then, for each vignette trial, participants first read were (1) presented with the background and vignette containing evidence and questions (but no sliders for inputs), as shown in the example interface in Figure 7 (**top**), and required to think about the vignette for 15 seconds without progressing; they coudl then could proceed to (2) an interface presenting sliders for the multi-click inputs, shown in Figure 7 (**bottom**). Participants took a median time of 2.24 minutes to provide all judgments for one vignette. Participants were paid at a base rate of $15/hr and told they may receive a bonus of up to $16/hr "if you try your best throughout the experiment to answer each question"; in reality, all participants were provided the bonus.

A4.4 Exp. 2: human judgment experimental details

This experiment followed the same interface format as Exp. 1, except with the underspecified backgrounds for each vignette. Additionally, participants in Exp. 2 were not shown the additional full description of the sport background prior to beginning the vignettes (they only read the backgrounds alongside the vignettes themselves). Participants took on average 2.81 minutes to provide their multi-click judgments per vignette. As with Exp. 1, participants in 2 of the 4 sets of vignettes also saw vignettes about diving, along with the other three sports, which were later omitted from the experimental analysis when this sport was removed from analysis.

A4.5 Exp. 3: human judgment experimental details

This experiment involved both a **human commentary elicitation** experiment and a **human judgment experiment**.

During the **commentary elicitation experiment**, as described in the main text, N=20 participants were shown a tutorial indicating that they would read vignettes about sports scenario, and then "act as a sports commentator" to write one or a few sentences introducing a new detail that would *change their* reasoning about a randomly selected new match prediction question. Participants were randomly assigned to conditions over which of the two new match questions they would need to change, and whether they were asked to produce details that would either *increase* or *decrease* the odds of a particular outcome given their initial judgments. On each trial, as in earlier experiments, participants completed the full judgment task – they read the vignettes for 15 seconds, then proceeded to the sliders where they entered judgments for all questions. They were then shown which new match prediction they were to manipulate and told the direction they would need to manipulate with their commentary. After writing commentary, participants were shown the full vignette (with their added commentary) and asked to re-enter their judgments on the new match prediction.

As noted in the main text, this experiment used only the tug-of-war and canoe racing sports; and used 9 base vignettes with the *underspecified* backgrounds from Exp. 2, and slightly different patterns of evidence than used in Exp. 1 and Exp. 2 – in particular, the vignettes included slightly easier outcome patterns of evidence involving head-head matches between single players, whereas the vignettes in Exp. 1 and Exp. 2 only involved matches between teams of two players each. Additionally, participants were shown slightly different wordings of the *judgment* questions during the trials: the strength questions asked *how strong* athletes were (rather than their absolute strength ranking out of random athletes) and *how much effort* they put in (rather than asking specifically for a percent effort).

In total, this experiment yielded an initial set of *81* initial distinct commentary observations across all participants. We filtered these down to 9 final vignettes by (1) excluding all participants who did not adjust their judgments after providing commentary in the specified direction (e.g., they did not actually increase the odds of the predicted match); (2) excluding participants who appeared to have used language models (participants were explicitly instructed not to) or who provided clearly spam answers; (3) excluding commentary that was more than a single sentence. We then selected the 9 commentary with a more specific set of criteria that could be generalized in future work – we selected commentary that focused on a *single athlete* (rather than generics about the world, like *it*

was raining); and commentary that focused on a single new *event* observation (eg. *Athlete A took an energy drink*) or observation about the athlete (*Athlete A had less experience*).

During the **human judgment experiment**, we then recruit a new set of participants to provide the same $k = 5$ multi-click judgments as in Experiments 1 and 2. Instructions were the same as in Exp. 1 and Exp. 2, except that participants were told that they would be reading vignettes including commentary written by other people. Each participant in this trial was shown the full set of k=9 vignettes with commentary. Participants were provided the *underspecified* sport description from Exp. 2, as described in the main text. Participants took approximately 2.22 minutes to provide their judgments per vignette.

A5 Results: Supplemental Analyses

This section collects additional analyses comparing human and model judgments.

A5.1 Human and MSA correlations between Experiments 1 and 2

We first examine how well *human judgments correlate across the matched vignettes in experiment 1 and experiment 2* – that is, whether people made judgments when reading the detailed background information that correlated with those from the underspecified background information. In general, as seen in Figure 8, judgments appear to be highly correlated, providing some evidence that people retrieve and use similar kinds of information to reason about the underspecified Exp. 2 condition as those that were provided to them explicitly in Exp. 1. Notably, there appears to be less correlation in the canoe race sport (middle column) – suggesting that people generally retrieved other ways that effort and strength might have contributed to the observed outcomes when left to come up with these details on their own, compared to the version spelled out to them in Exp. 1.

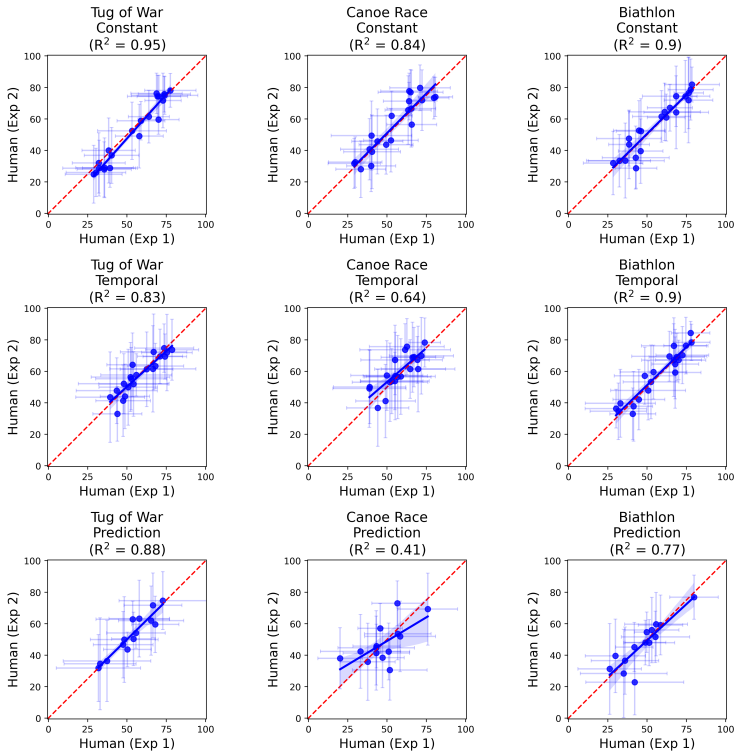

Figure 8: Correlations between human participant predictions per stimuli per query between Experiment 1 (x axis) versus Experiment 2 (y axis).

We perform the same analysis for the MSA judgments, comparing correlations between MSA judgments in Experiment 1 and 2 (Figure 9). In general, we see that the judgments are *very* well

correlated between the experiments, more so than the human participants – and unlike human participants, we do not see variance in the canoe racing condition between Experiment 1 and 2. This warrants further investigation, as it suggests that the model synthesis procedure is less diverse in producing human-like distributions over possible ad-hoc *models* in Experiment 2 from underspecified backgrounds, and may reflect a lack of sampling diversity in model construction.

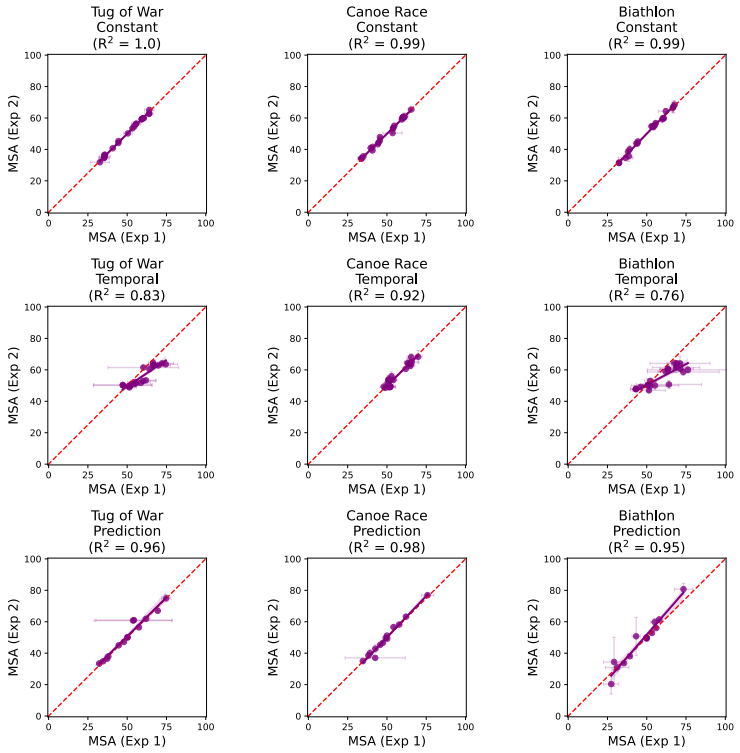

Figure 9: Comparing correlations between MSA predictions per stimuli per query for Experiment 1 (x axis) versus Experiment 2 (y axis).

A5.2 Total Variation Distance for comparing distributions between humans and models

To ensure our distributional analyses are not specific to using the Wasserstein Distance metric, we repeat our distributional analyses using Total Variation Distance (which does not account for the "geography" of the domain when comparing distributions). As in our Wasserstein Distance analyses, we first bucketize participant and model judgments (into 10 buckets) and compute our measure over the buckets. We see similar trends across models, sports, and experiments in Figure 10.

A5.3 Human-Model Correlations for All Models

Below, we include the full set of scatterplots between the average human and average model responses for the three experiments. We depict additional gold model results for Exp. 1 and Exp. 2 in Figure 11, Direct-LLM in Figure 12, and CoT-LLM in Figure 13. We compare all model scatterplots on Exp. 3 in Figure 14.

A5.4 Human-Model Correlations for All Models

Figure 15 also briefly summarizes qualitative error analysis patterns between Experiments 1 and 2, highlighting distinctions in LM-only baselines relative to human judgments in overall patterns of judgments (red) – as well as distinctions between the MSA baselines in the *qualitative* nature of the distributions of human judgments (LMs often appear to make peakier judgments relative to the symbolic model posteriors, which could be an artifact of the 5-sampling procedure).

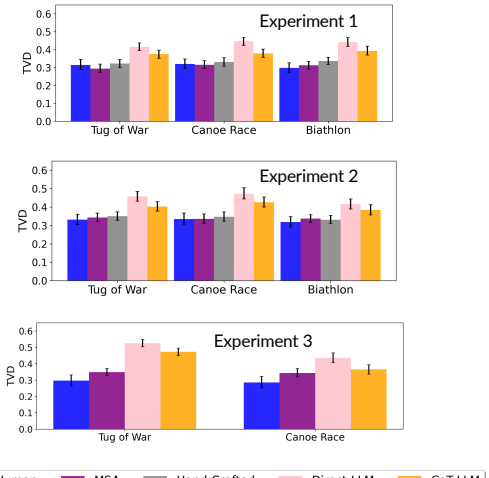

Figure 10: Comparing Total Variation Distance (TVD) between model and human judgments across each experiment. Bootstrapping and averaging follow as in our Wasserstein Distance computations; that is: TVD is computed between judgments per query per scenario, then aggregated as the mean over query types, and mean across query types for each depicted sport and experiment.) Error bars for model-humans show 95% CI over 1000 bootstrapped samples, with replacement, on the human data; for human-humans, over 1000 sampled 50-50 split-halve TVDs.

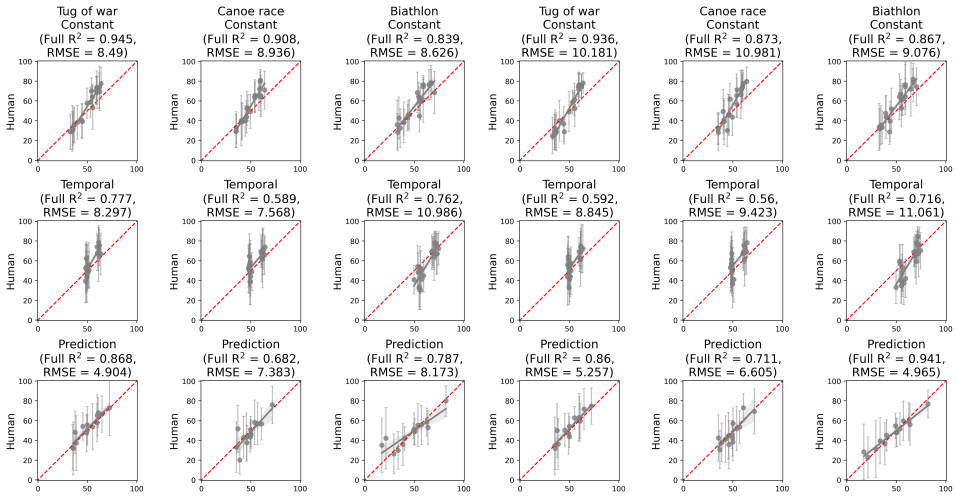

Figure 11: Inferences under the gold model against people for Exp. 1 (left) and Exp. 2 (right). Error bars depict standard deviation over the human responses.

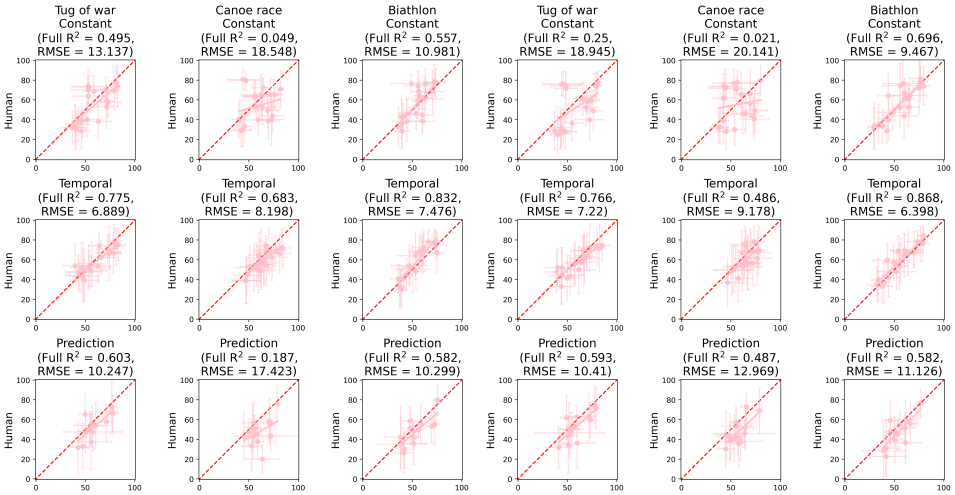

Figure 12: Inferences under the Direct-LLM model against people for Exp. 1 (left) and Exp. 2 (right). Error bars depict standard deviation over the human and model responses.

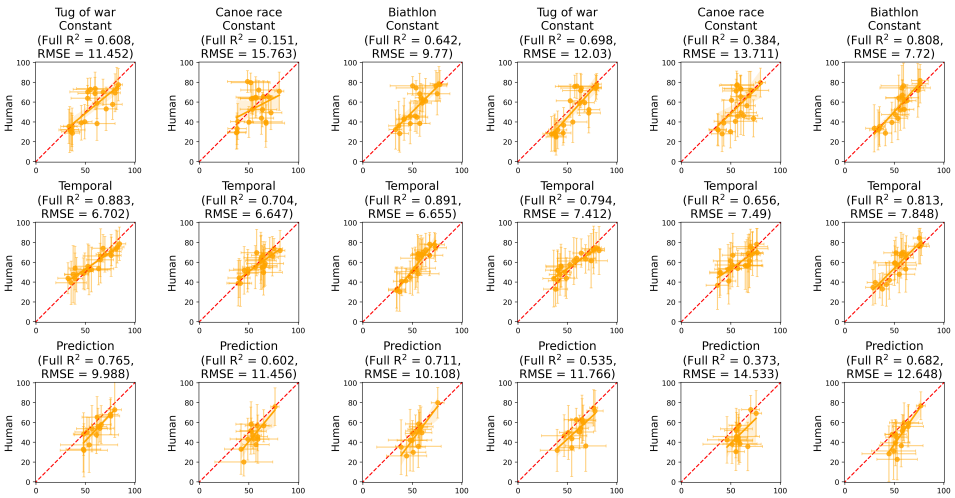

Figure 13: Inferences under the CoT-LLM model against people for Exp. 1 (left) and Exp. 2 (right). Error bars depict standard deviation over the human and model responses.

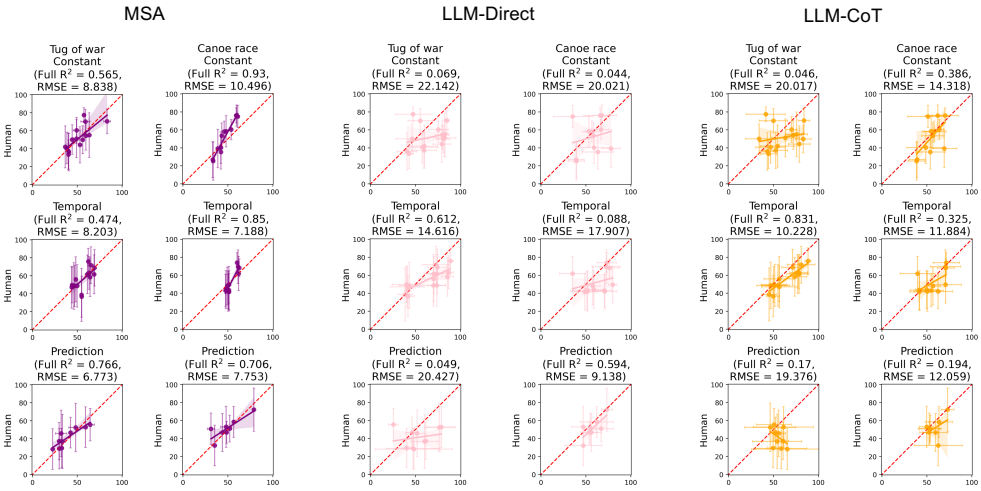

Figure 14: Inferences under MSA (left), Direct-LLM (middle), and CoT-LLM (right) against people for Exp. 3. Error bars depict standard deviation over the human and model responses.

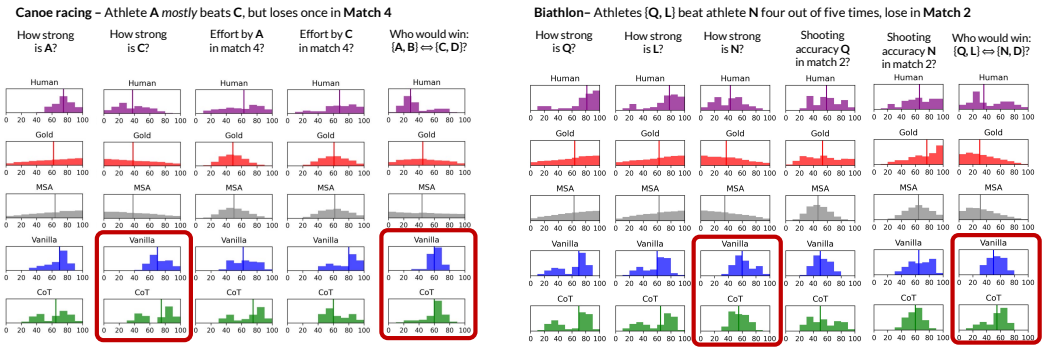

Figure 15: Illustrative examples from Experiments 1 and 2 highlighting one divergent pattern in inferences from LLM-only baselines, relative to human judgments (and normative Bayesian inferences in models synthesized by our MSA implementation). In the *canoe racing scenario* (**left**), noisy evidence in the vignette suggests that athlete *A* often appears in a winning pair of teammates that beat teams containing athlete *C*, despite a single anomalous loss. Humans judge C to be largely weaker than average, but both LLM-baselines (red) switch to predicting C as particularly strong; and predict that C would now win on a team gainst A. Similarly, in the *biathlon scenario* (**right**), a *pair* of athletes (Q, L) frequently beats another (N), while losing anomalously once. LLM-baselines allocate more probability to the possibility that both Q and L are actually quite weak, largely believe N is stronger, and tend to predict that N will win against Q and L.