EMERGENT BAYESIAN BEHAVIOUR AND OPTIMAL CUE COMBINATION IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) excel at explicit reasoning, but their implicit computational strategies remain underexplored. Decades of psychophysics research show that humans intuitively process and integrate noisy signals using near-optimal Bayesian strategies in perceptual tasks. We ask whether LLMs exhibit similar behaviour and perform optimal multimodal integration without explicit training or instruction. Adopting the psychophysics paradigm, we infer computational principles of LLMs from systematic behavioural studies. We introduce a behavioural benchmark - BayesBench: four magnitude estimation tasks (length, location, distance, and duration) over text and image, inspired by classic psychophysics, and evaluate a diverse set of nine LLMs alongside human judgdments for calibration. Through controlled ablations of noise, context, and instruction prompts, we measure performance, behaviour and efficiency in multimodal cue-combination. Beyond accuracy and efficiency metrics, we introduce a Bayesian Consistency Score that detects Bayes-consistent behavioural shifts even when accuracy saturates. Our results show that high task accuracy - notably for GPT-5 Mini - does not always imply efficient cue combination; yet accurate models, including GPT-5 Mini, Llama-4 Maverick, and Claude 3.7 Sonnet, often adapt in Bayes-consistent ways. These findings reveal emergent principled handling of uncertainty and highlight the correlation between accuracy and Bayesian tendencies. We release our psychophysics benchmark and consistency metric as evaluation tools and to inform future multimodal architecture designs.

1 INTRODUCTION

The estimation of magnitudes, including quantities like length, duration, or distance, represents one of the most fundamental computations in biological and artificial intelligence. Humans perform these judgdments through the Bayesian integration of noisy sensory signals, automatically weighting cues by their reliability (Ernst & Banks, 2002) and incorporating prior expectations to minimise estimation error (Remington et al., 2018; Knill & Pouget, 2004). This computational strategy emerges without explicit instruction across diverse cultures and developmental stages, suggesting it reflects a fundamental solution to information processing under uncertainty.

This universality raises the critical question of whether modern LLMs, trained solely on next-token prediction without explicit perceptual objectives (Radford et al., 2018), spontaneously develop analogous computational strategies. Understanding how LLMs process and integrate uncertain information has immediate implications for building robust multimodal systems that appropriately handle varying input quality (Kendall & Gal, 2017; Ma et al., 2022). Moreover, if large computational models naturally converge on Bayesian principles for handling uncertainty, it would suggest that these principles emerge from information-theoretic constraints rather than biological evolution—implying a deeper computational universality (Barlow et al., 1961; Wei & Stocker, 2015).

To investigate this, we apply classical psychophysics methodology (Petzschner et al., 2015) to probe these implicit computational strategies in LLMs, treating them as black-box observers and inferring their mechanisms from systematic behavioural analysis. By controlling stimulus uncertainty and measuring characteristic signatures of Bayesian processing, we can determine whether LLMs exhibit human-like optimal perception without explicit training. As a result, we present three contributions: 1) We introduce a systematic psychophysics framework for LLMs, a reproducible pipeline for four

synthetic magnitude estimation tasks probing length, location, distance, and duration. Our pipeline allows controlled ablations of noise, context, and instruction prompts to track behavioural changes. 2) We develop a new benchmark: BayesBench based on task performance, cue-combination efficiency, and Bayesian consistency computed with a novel Bayesian Consistency Score. 3) We show evidence of emergent Bayes-consistent behaviour in capable LLMs and its correlation with task performance.

2 Related Work

Human psychophysics. The quantitative study of perception has revealed systematic relationships between physical stimuli and perceptual judgements, formalised in classical laws like Weber-Fechner's logarithmic scaling and Vierordt's temporal regression effects (Fechner, 1860; Weber, 1834; Gibbon, 1977; Jazayeri & Shadlen, 2010; Roseboom et al., 2019; Fountas & Zakharov, 2023). These phenomena, including scalar variability and sequential biases, emerge from optimal Bayesian inference under uncertainty (Petzschner & Glasauer, 2011). When observers estimate magnitudes, they automatically combine noisy measurements with prior expectations, producing characteristic behavioural patterns. Figure 1 illustrates this regression-to-the-mean effect in both Llama-4 Maverick's responses and human psychophysics data—evidence of shared computational principles despite vastly different substrates, as we will see in later sections.

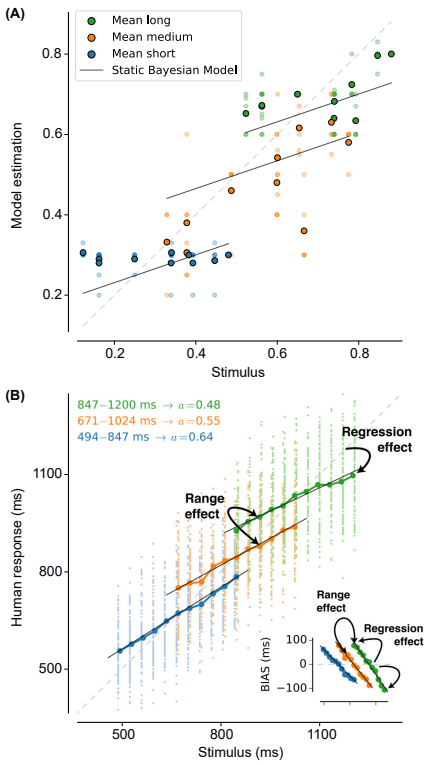

Figure 1: Comparison of LLMs vs human behaviour **A:** Response of Llama-4 Maverick in one of the line length ratio estimation experiments. The fitted lines are based on a static Bayesian observer model. **B:** Response from typical human psychophysics studies (adapted from Thurley, 2016). We see in both that there is a regression to the mean effect, where responses are biased towards the centre of the stimulus range.

LLMs and Bayesian behaviour. Certain aspects of LLMs are shown to be consistent with Bayesian computation. For example, in-context learning can be interpreted as approximate Bayesian inference (Xie et al., 2021) and, in reasoning, Bayesian teaching is shown to improve performance (Qiu et al., 2025). Similarly, LLMs spontaneously segment sequences using Bayesian surprise in ways that correlate with human event perception (Fountas et al., 2025). However, most studies probe explicit reasoning or learned behaviours, where models can leverage acquired statistical rules, rather than perceptual tasks that could reveal computational strategies emerging implicitly from pretraining.

Multimodal studies. Progress have been rapid in developing multimodal LLMs, alongside this is the deployment of benchmarks such as MMbench (Liu et al., 2024) and SEED-bench (Li et al., 2024) that test multimodal reasoning. However, most of these benchmarks do not cover controlled manipulations of modality specific noise for studying fusion strategies. Our synthetic datasets allow fine-grained cue-combination analysis and studies how LLMs combine noisy information from multiple modalities. This is still a nascent area of research but crucial for better understanding how we may build more robust and generalisable models that will behave optimally under uncertainty.

3 Methods

3.1 Estimation tasks and ablations

We develop four psychophysics-inspired magnitude–estimation tasks illustrated in Figure 2:

- **Marker location estimation:** given a line with a red marker (or '0' in text input) estimate the position of a marker on a line as a number between 0 to 1.
- **Line ratio estimation:** given two lines, estimate the ratio of the shorter line to the longer line.
- **Maze distance estimation:** given a non-self-intersecting path, estimate the straight line distance between start and the end of the path.
- **Duration estimation:** given an extract of a conversation transcript, estimate the duration of the dialogue. Transcripts are extracted from the AMI Meeting Corpus Kraaij et al. (2005).

The first three tasks are multimodal, with text input and image inputs.

We conduct ablations to probe LLMs and analyse changes in behaviours (see Appendix A.3 for ablation details):

- **Steering:** provide additional textual or numerical information in the system prompt. Aimed at studying how LLMs behaviour changes when asked to consider uncertainty in its responses.
- **Noise:** add constant or gradually increasing blur to the image modality. Aimed at studying how LLMs may reweight information in the presence of noise.
- **Context:** change the length of the available history or reversing trial sequence. Aimed at studying how previous context affects behaviour.

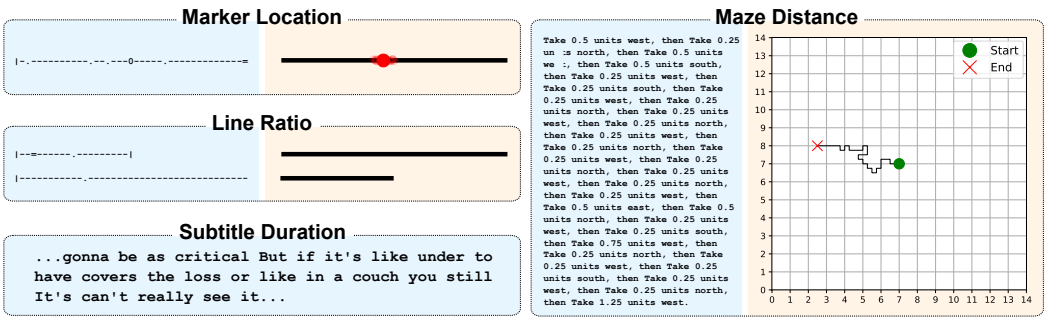

Figure 2: Example of the four magnitude estimation tasks. Cues in a blue background represent information provided as text, while orange represents vision.

3.2 Behavioural Modelling

In human psychophysics studies, participants' responses are fitted against a range of behavioural models to infer their internal computational strategies (Petzschner & Glasauer, 2011; Jazayeri & Shadlen, 2010). This is an effective approach when the subject is essentially a black box, and we can only observe their input-output behaviour. In line with this framework, we fit LLMs' responses against a set of behavioural models covering factors of interest. The degree of fit to different models indicates the extent to which LLMs exhibit that behaviour.

In the below, x_t and y_t denote the true input value of the stimulus and the LLM's estimate at trial t, respectively. μ_t and $\sigma_{\text{dec}}$are the LLM's internal estimate and response noise level, respectively. We used three main types of behaviour models:

Linear observer. LLM's estimation of the input stimuli is a linear function of the stimulus value:

$$\mu_t = wx_t + b, \quad y_t \sim \mathcal{N}(\mu_t, \sigma^2_{\text{dec}}). \tag{1}$$

Static Bayesian observer. LLM's estimation is a weighted average of the input stimulus x_t and a fixed prior belief μ_p:

$$\mu_t = \frac{\tau_x}{\tau_x + \tau_p} x_t + \frac{\tau_p}{\tau_x + \tau_p} \mu_p, \quad y_t \sim \mathcal{N}(\mu_t, \sigma^2_{\text{dec}}), \tag{2}$$

τ_x and τ_p denote the measurement and prior precisions respectively. We show in the upper panel of Figure 1 an example where this model best fits the LLM's responses.

Sequential Bayesian observer (Kalman filter). LLM's estimation is updated trial-by-trial following a standard Kalman filter:

$$\mu_{t|t-1} = \mu_{t-1|t-1}, \quad P_{t|t-1} = P_{t-1|t-1} + q, \quad y_t \sim \mathcal{N}(\mu_t, \sigma^2_{\text{dec}}), \tag{3}$$

Where the update equations are:

$$K_t = \frac{P_{t|t-1}}{P_{t|t-1} + r}, \quad \mu_{t|t} = \mu_{t|t-1} + K_t(x_t - \mu_{t|t-1}), \quad P_{t|t} = (1 - K_t)P_{t|t-1}. \tag{4}$$

r is the measurement noise variance, q is the process noise variance and P is the variance about its estimate. We show in Figure 3 and 4 an example where this model best fits the LLM's responses.

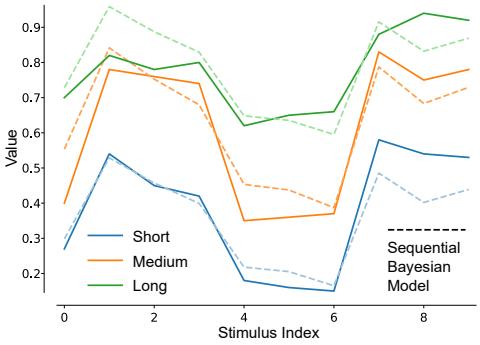

Figure 3: GPT-5 Mini's mean response (to verbal cues) compared to prediction based on a sequential Bayes model (dotted line)

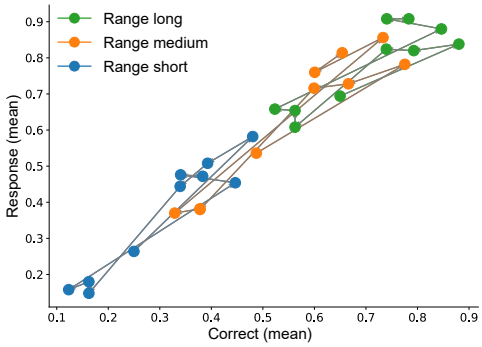

Figure 4: GPT-5 Mini's mean response trajectory (verbal cue). Arrows denote the sequence of its responses.

Additional model variants Across all models, we include variants with an additional stage of log-transform on the input and output (this mimics studies that support evidence of a logarithmic perception of magnitude in humans and animals (Nieder & Miller, 2003; Nover et al., 2005).)

For non-linear models, we fitted variants where a final stage of gain or affine transformation is applied. This is to account for potential mis-calibration in output mapping (this is not needed for linear models as it is captured in the bias and gradients). Further details can be found in Appendix A.11.

3.3 Cue combination modelling

For our multimodal tasks, we study how LLMs combine text and image cues by modelling their multimodal responses against their unimodal responses. The main models are in Table 1. y_{comb}, y_{text} and y_{image} denote the LLM's response for multimodal, unimodal text and unimodal image, respectively.

Equal weighting	**Linear regression**	**Bayes-optimal fusion**
$y_{\text{comb}} = \frac{1}{2}(y_{\text{text}} + y_{\text{image}})$	$y_{\text{comb}} = \alpha\, y_{\text{text}} + (1-\alpha)\, y_{\text{image}}$	$y_{\text{comb}} = w_{\text{text}}\, y_{\text{text}} + \big(1 - w_{\text{text}}\big)\, y_{\text{image}}$ $w_{\text{text}} = \dfrac{1/\sigma^2_{\text{text}}}{1/\sigma^2_{\text{text}} + 1/\sigma^2_{\text{image}}}$

Table 1: Cue-combination baselines. α is fitted in $[0, 1]$. σ^2_{text} and σ^2_{image} are the empirical variances of the LLM's responses in the text-only and image-only conditions respectively.

For the *Bayes-optimal fusion* model, we report *Oracle* (calibrated, covariance-based) and *Non-Oracle* (uncalibrated, variance-based) variants. This fusion is the optimal *linear unbiased* combiner (BLUE) under linear-Gaussian assumptions. See Appendix A.6 for details.

3.4 Model evidence

We compare model fit for behavioural modelling and cue-combination modelling based on Akaike Information Criterion (AIC). Different behavioural model variants differ along interpretable factors

Factor	Expression	Parameters
NRMSE (A)	$1 - \frac{\text{NRMSE} - \text{NRMSE}_{\min}}{\text{NRMSE}_{\max} - \text{NRMSE}_{\min}}$	$\text{NRMSE}_{\min,\max} = 0,\ 2$
RRE (E)	$\big[\text{RRE}(\text{Bayes Oracle}) + \text{RRE}(\text{Bayes Non Oracle})\big]/2$	N/A
BCS (C)	$(\text{BCS} - \text{BCS}_{\min})/(\text{BCS}_{\max} - \text{BCS}_{\min})$	$\text{BCS}_{\min,\max} = -15,\ 15$

Table 2: BayesBench components. $\text{NRMSE}_{\max}$ is set to 2 (twice the error committed by the constant predictor baseline). $\text{BCS}_{\min,\max}$ are set equal to the range of scores for five ablations across three multimodal experiments.

(e.g. Bayesian vs Non-Bayesian or Sequential vs Non-Sequential). We focus on the Bayesian vs Non-Bayesian dimension and compute factor evidence by comparing the best fitting models in each category. See Appendix A.7 for further details.

3.5 Key metrics

We quantify (i) task accuracy, (ii) cue–combination efficiency, and (iii) behavioural consistency.

Accuracy (NRMSE). $\text{NRMSE} = \text{RMSE}_{\text{model}}/\text{RMSE}_{\text{baseline}}$, where RMSE has the standard root-mean-squared-error definition. The baseline is a constant predictor that outputs the mean of the stimulus range (lower is better).

Efficiency (RRE). $\text{RRE}(m_{\text{ref}}) = \text{NRMSE}_{\text{ref}}/\text{NRMSE}_{\text{LLM}}$ for any reference combiner m_{ref} (Sec. 3.3). RRE values > 1 (< 1) mean the LLM has lower (higher) error than the reference.

Bayesian Consistency Score (BCS). To test whether LLM's behaviour shifts in the *Bayes–consistent* direction under controlled ablations, we compare the fitted weights of a static Bayesian observer model (Sec. 3.2). The posterior mean of this model is precision–weighted with $w_{\text{prior}} = \tau_p/(\tau_p + \tau_x)$ (prior precision τ_p, measurement precision τ_x), so increasing τ_p or decreasing τ_x raises w_{prior}.

We use five ablations across three tasks to compute BCS. These ablations are designed to increase τ_p and/or decrease τ_x: (i) **Steering (verbal)** and (ii) **Steering (unbiased numerical)** provide range–consistent context or prompt the model about measurement noise, effectively strengthening the prior ($\tau_p \uparrow$); (iii) **Noise (constant)** and (iv) **Noise (gradual)** blur image inputs to reduce measurement precision ($\tau_x \downarrow$); (v) **Context (longer context window)** supplies a longer rolling history without altering current measurements ($\tau_p \uparrow$, τ_x unchanged).

For each ablation a, we compare fitted weights to the base experiment, $\Delta w_{\text{prior}} = w_{\text{prior}}^{(\text{ablation})} - w_{\text{prior}}^{(\text{base})}$, and set

$$s_a = \begin{cases} +1 & \text{if } \Delta w_{\text{prior}} \geq 0, \\ -1 & \text{if } \Delta w_{\text{prior}} < 0, \end{cases} \qquad \text{with } s_a = 0 \text{ if } w_{\text{prior}}^{(\text{ablation})} > 0.9.$$

We focus on the *sign* of Δw_{prior} since magnitudes depend on model–specific factors, such as how accurate a given model's perception is. For example, a highly perceptually accurate model may only need to adjust w_{prior} by a smaller amount given an injection of measurement noise. We set s_a to zero when $w_{\text{prior}}^{(\text{ablation})} > 0.9$, because this indicates a prior-dominant regime, where the model is essentially disregarding the current stimulus and always outputting a constant. This is undesirable because in all five selected ablations the stimulus should remain informative.

The *Bayesian consistency score* sums over ablations: $\text{BCS} = \sum_a s_a$.

3.6 Composite Benchmark Score (BayesBench).

The overall *BayesBench score* is a function of three metrics: NRMSE factor for task accuracy (A), RRE factor for cue-combination performance against a Bayes-optimal reference (E) and BCS factor

for Bayes-consistency in behaviour adaptation (C) (defined in Table 2). The first factor is averaged across all four tasks while the latter two are averaged across the three multimodal tasks. In the (A) factor, $\mathrm{NRMSE}_{\max} = 2$ defines the upper bound of model NRMSE and models that incur larger error receive no credit. This range spans our model range and marks the worst reasonable performance of any model.

The *BayesBench score* is defined as:

$$S_{\text{BayesBench}} = \tfrac{1}{3}\left(A + E + C\right). \tag{5}$$

4 Experimental Setup

Each estimation task is divided into three sessions (short, medium, long), where stimulus values fall in different but overlapping ranges per session. This overlap allows us to study context-dependent effects. Figure 5 provides an overview of the stimulus value distributions across sessions, using the marker location as an example. In each session, at each trial, the LLM is given the context of its prior trials (i.e., both the stimulus probes and the LLM's previous responses, as each API interaction is statless or "memoryless"). The rolling context simulates how humans form memory of recent interactions, and is the basis of the emergence of Bayesian consistent behaviour. The overall view of our experimental setup is shown in Appendix A.1.

Interactions with LLMs are performed via API. See Appendix A.2 for further details.

We evaluate a diverse set of recent LLMs spanning closed- and open-weight releases (see Appendix A.4 for details). Where possible, we disable extended-thinking or reasoning controls to probe the models' natural, emergent behaviour. This was feasible for all models except GPT-5 mini, which only allows adjusting reasoning depth; we set this to the lowest level.

In addition, we ran a human baseline study for comparison on all our tasks under a small number of ablations. See Appendix A.5 for details. Human results are included in the left panel of Figure 7. This experiment and analysis are used only as a reference point here, as our main focus is on comparing LLMs against each other. Extensive human psychophysics studies, including the magnitude estimation effects examined here, are extensively documented in psychophysics literature, such as in Jazayeri & Shadlen (2010); Petzschner & Glasauer (2011).

We estimate uncertainty in our results using 30 rounds of bootstrapping while preserving the trial structure within each session to maintain contextual integrity. The error bars shown in Figure 7 and 8 represent 68% bootstrap percentile intervals.

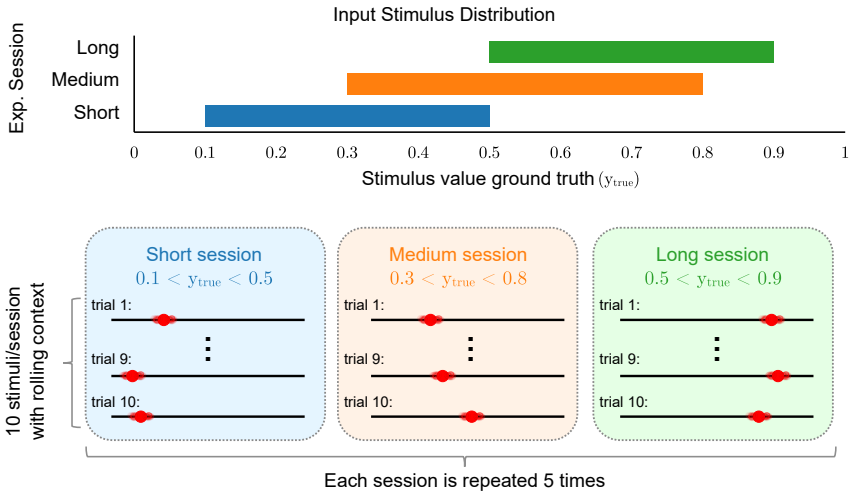

Figure 5: Example distribution of stimulus input for the marker location task

5 RESULTS

5.1 OVERALL PERFORMANCE AND BEHAVIOURAL FIT

Most models perform better in the *text* than the *image* modality, except on the maze distance estimation task. The text input for the maze distance estimation task is a long, detailed path description—much more complex than the ASCII prompts used in other tasks (Fig. 2). Most models perform worse in the text modality for this task, but GPT-5 Mini is an outlier here: it achieves near-perfect text performance, likely due to residual reasoning that we can attenuate but not fully disable. Across tasks, the factor evidence for Bayesian behaviour is consistently higher in the image modality than in text (Appendix A.9).

Moving from unimodal to multimodal inputs does not uniformly improve performance. However, some models are better able to leverage information from the additional modality: Llama-4 Maverick attains its best performance under multimodal conditions across all tasks, and Claude 3.7 Sonnet and GPT-4o improve on two of the three tasks.

Overall, the strongest models (GPT-5 Mini, Claude 3.7 Sonnet, GPT-4o) reach low error rates, comparable to—or better than—human performance (left panel, Fig. 7).

With the exception of Gemini 2.5 Flash Lite, there is a general trend in the left panel of Figure 7 that more accurate models also show stronger evidence of Bayesian behaviour.

5.2 CUE COMBINATION

From the middle panel of Figure 7, we see that not all models with good NRMSE performance also exhibit efficient cue combination. GPT-5 Mini, despite its strong NRMSE performance, shows poor cue combination efficiency. This is especially pronounced in the maze distance estimation task, where GPT-5 Mini's performance in the text modality is essentially perfect and much better than its image modality performance. This implies that a Bayes-optimal combination must significantly further downweight its image input. However, it appears unable able to downweight its image input to the optimal extent (see Appendix A.10 for further details).

On the other hand, Llama-4 Maverick's multimodal NRMSE performance exceeds that of a Bayesian reliability-weighted unbiased linear combination. In Figure 6, we fitted Llama-4 Maverick's multimodal responses against its unimodal responses. We found that a random forest is able to fit its multimodal responses from unimodal responses better than either the Bayes Non Oracle model or a linear regression model. This suggests that Llama-4 Maverick is likely using a more sophisticated, non-linear, cue combination strategy.

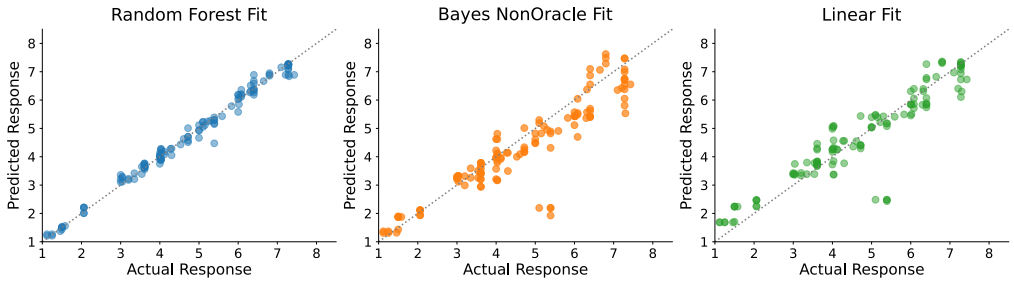

Figure 6: Comparison of cue combination model fits for Llama-4 Maverick. Left panel: random forest fit (blue). Middle panel Bayes-optimal fit (orange). Right panel: linear regression fit (orange).

5.3 BAYESIAN CONSISTENCY

From the right panel of Figure 7, we see that generally more accurate models also tend to exhibit more Bayes-consistent behaviour. While Gemma 3 4B and Phi 4 Multimodal achieved higher BCS than expected, they are also the least accurate group of models.

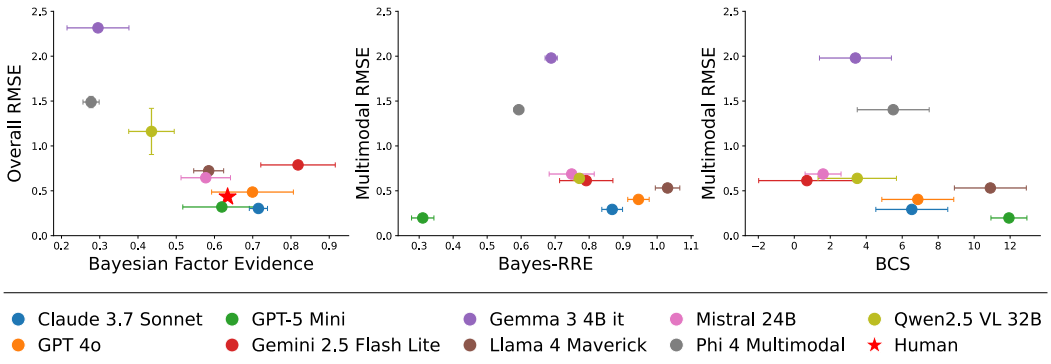

Figure 7: Results summary across models and tasks. Left panel: Bayesian behavioural evidence and relationship against overall NRMSE. Middle panel: cue-combination performance. Shows relationship between multimodal tasks NRMSE and efficiency against Bayes-optimal cue-combination reference models. Right panel: Bayes-consistency score and its relationship against multimodal NRMSE. Each point represents a model, with color indicating model family. Error bars represent 68% bootstrap percentile intervals. Human baseline is shown in the left panel for reference.

5.4 BayesBench Summary

Figure 8 shows the computed BayesBench scores across models, in accordance to the definition in Section 3.6. Bayes-RRE generally increases with accuracy (lower NRMSE), with two notable exceptions: GPT-5 Mini underperforms on Bayes-RRE relative to its NRMSE, whereas Llama-4 Maverick exceeds expectations on Bayes-RRE. BCS likewise tends to track accuracy but provides additional separation among the top models. Overall, Llama-4 Maverick attains the highest BayesBench score, driven by strong Bayes-RRE and BCS components.

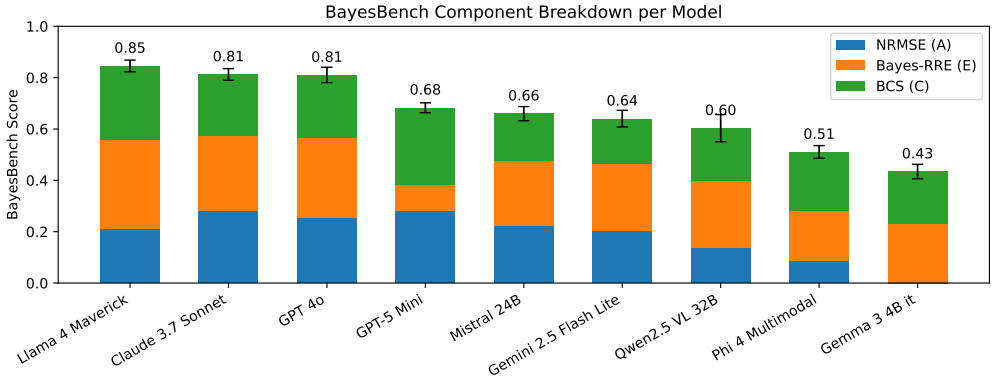

Figure 8: BayesBench overall score, with breakdown into components. Error bars represent 68% bootstrap percentile intervals.

6 Discussion

Our study reveals that LLMs exhibit rich and diverse behavioural patterns when probed with psychophysics-inspired magnitude estimation tasks. While the degree of factor evidence for Bayesian behaviour differs by task and modality, more accurate models (e.g., GPT-5 Mini, Claude 3.7 Sonnet, Llama-4 Maverick) tend to display higher Bayesian factor evidence, especially in the image modality (Appendix A.9 for full breakdown). These models tend to adapt their behaviour in Bayes-consistent ways when inputs are subjected to perturbations such as noise, steering, or extended context (right panel of Figure 7). This supports the view that LLMs behave in ways consistent

with approximate Bayesian observers, even without explicit training or reasoning instructions. This is reminiscent of findings in human psychophysics, where Bayesian models explain a wide range of human perceptual phenomena and processing in the brains (Petzschner et al., 2015; Knill & Pouget, 2004) without explicit training.

We find that high task accuracy does not always imply optimal cue–combination (middle panel of Figure 7). For example, GPT-5 Mini attains very low NRMSE yet does not combine modalities efficiently compared to other models. This shortfall is most apparent when unimodal performance is imbalanced: optimal behaviour would require the model to markedly down-weight the weaker modality, which some LLMs fail to do. Conversely, Llama-4 Maverick surpasses Bayesian reliability-weighted linear fusion, indicating the use of more sophisticated non-linear integration strategies.

Comparing uni- and multimodal performance reveals that, while models such as Llama-4 Maverick, Claude 3.7 Sonnet, and GPT-4o are able to utilise the additional modality of input to achieve lower error when both modalities (text and image) are present for the the majority of multimodal tasks, this is not a universal trend. The variability in gains indicates potential headroom for advancing multimodal LLMs. See Appendix A.9 for model-specific breakdown.

To capture behavioural features beyond static task metrics, we devised the Bayesian Consistency Score (BCS) that captures principled behavioural shifts. This allows us to evaluate model behaviour more holistically, even when accuracy saturates. Measuring behaviour changes under controlled ablations enable us to compare models that may have different base performance and can offer additional insights into implicit computational strategies.

Our results show that LLMs are generally consistent with Bayesian observer models. This raises the question of how Bayesian computation can be an emergent property of sufficiently capable models trained on large-scale data, similar to questions tackled in human studies (Barlow et al., 1961; Wei & Stocker, 2015). Future architectures or training regimes that better encode uncertainty and support principled cue combination may improve LLMs' robustness in noisy, real-world settings. Furthermore, benchmarks such as our custom BayesBench can complement standard accuracy-based evaluations, offering diagnostic insights into implicit computational strategies.

Limitations. Our tasks are synthetic and designed for precise control. It remains an open question how well the observed behaviours generalise to naturalistic multimodal environments. As the test range of our tasks is bounded, effects that only emerge with longer sequences may not be detected. Our ablation studies are necessarily limited in scope; other perturbations, such as different noise types, may illustrate different aspects of behaviour. In addition, all interactions relied on API access, which may be affected by API non-determinism or silent vendor updates.

7 Conclusion and Future Directions

We present BayesBench, a psychophysics-inspired benchmark that probes LLMs' ability to estimate magnitudes, integrate noisy multimodal cues, and exhibit Bayes-consistent behaviour. Our findings show that capable LLMs not only achieve low error rates but also adapt in Bayesian consistent manners, revealing emergent cognitive-like strategies. Strong multimodal models can also combine cues efficiently, although this is not guaranteed by high accuracy alone. Our results suggest that Bayesian-consistent behaviour may emerge naturally in sufficiently capable models.

Our work bridges human psychophysics and AI research, by providing both an extensible template and a set of diagnostic metrics. While our tasks are synthetic, they highlight possible directions for studying implicit computation in LLMs. Future work should extend the dataset to more naturalistic multimodal domains, explore representational underpinnings from a mechanistic perspective, and assess how Bayesian tendencies scale with model size and training data. It may also be fruitful to further develop more sophisticated diagnostic metrics that capture specific aspects of behavioural adaptations, extend ablation studies and enlarge the scope of human baselines for comparison.

Reproducibility Statement We will release *BayesBench* for public use, including the synthetic data generator, prompts, ablation configurations, behavioural/cue–combination model code and evaluation scripts. The behavioural models are fully specified in Section 3.2; cue–combination mod-

els are fully specified in Section 3.3; factor-evidence computation in Section 3.4 and Appendix A.7; the metrics and composite score is specified in Section 3.5 and Appendix A.8;

REFERENCES

Horace B Barlow et al. Possible principles underlying the transformation of sensory messages. *Sensory communication*, 1(01):217–233, 1961.

Marc O Ernst and Martin S Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433, 2002.

Gustav Theodor Fechner. *Elemente der Psychophysik*. Breitkopf und Härtel, Leipzig, 1860. English translation: *Elements of Psychophysics* (Adler, 1966), Holt.

Zafeirios Fountas and Alexey Zakharov. Bayesian sense of time in biological and artificial brains. In *TIME AND SCIENCE: Volume 2: Life Sciences*, pp. 237–265. World Scientific, 2023.

Zafeirios Fountas, Martin Benfeghoul, Adnan Oomerjee, Fenia Christopoulou, Gerasimos Lampouras, Haitham Bou Ammar, and Jun Wang. Human-inspired episodic memory for infinite context LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=BI2int5SAC.

John Gibbon. Scalar expectancy theory and weber's law in animal timing. *Psychological review*, 84 (3):279, 1977.

Mehrdad Jazayeri and Michael N Shadlen. Temporal context calibrates interval timing. *Nature neuroscience*, 13(8):1020–1026, 2010.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

David C Knill and Alexandre Pouget. The bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, 27(12):712–719, 2004.

Wessel Kraaij, Thomas Hain, Mike Lincoln, and Wilfried Post. The ami meeting corpus. In *Proc. International Conference on Methods and Techniques in Behavioral Research*, pp. 1–4, 2005.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13299–13308, 2024.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18177–18186, 2022.

Andreas Nieder and Earl K Miller. Coding of cognitive magnitude: Compressed scaling of numerical information in the primate prefrontal cortex. *Neuron*, 37(1):149–157, 2003.

Harris Nover, Charles H Anderson, and Gregory C DeAngelis. A logarithmic, scale-invariant representation of speed in macaque middle temporal area accounts for speed discrimination performance. *Journal of Neuroscience*, 25(43):10049–10060, 2005.

Frederike H Petzschner and Stefan Glasauer. Iterative bayesian estimation as an explanation for range and regression effects: a study on human path integration. *Journal of Neuroscience*, 31 (47):17220–17229, 2011.

Frederike H Petzschner, Stefan Glasauer, and Klaas E Stephan. A bayesian perspective on magnitude estimation. *Trends in cognitive sciences*, 19(5):285–293, 2015.

Linlu Qiu, Fei Sha, Kelsey Allen, Yoon Kim, Tal Linzen, and Sjoerd van Steenkiste. Bayesian teaching enables probabilistic reasoning in large language models. *arXiv preprint arXiv:2503.17523*, 2025.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Evan D Remington, Tiffany V Parks, and Mehrdad Jazayeri. Late bayesian inference in mental transformations. *Nature Communications*, 9(1):4419, 2018.

Warrick Roseboom, Zafeirios Fountas, Kyriacos Nikiforou, David Bhowmik, Murray Shanahan, and Anil K Seth. Activity in perceptual classification networks as a basis for human subjective time perception. *Nature communications*, 10(1):267, 2019.

Kay Thurley. Magnitude estimation with noisy integrators linked by an adaptive reference. *Frontiers in Integrative Neuroscience*, 10:6, 2016.

Ernst Heinrich Weber. *De pulsu, resorptione, auditu et tactu: Annotationes anatomicae et physiologicae*. C. F. Koehler, Leipzig, 1834.

Xue-Xin Wei and Alan A Stocker. A bayesian observer model constrained by efficient coding can explain'anti-bayesian'percepts. *Nature neuroscience*, 18(10):1509–1517, 2015.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021.

A Appendix

A.1 Experimental design

The basic setup of our experiments follows: 1) dataset and prompt generation, 2) session structure covering the order of stimulus presentation, 3) interation with LLM via API and 4) analyses. This modular design (Figure 9) allows for systematic exploration of different factors influencing model performance.

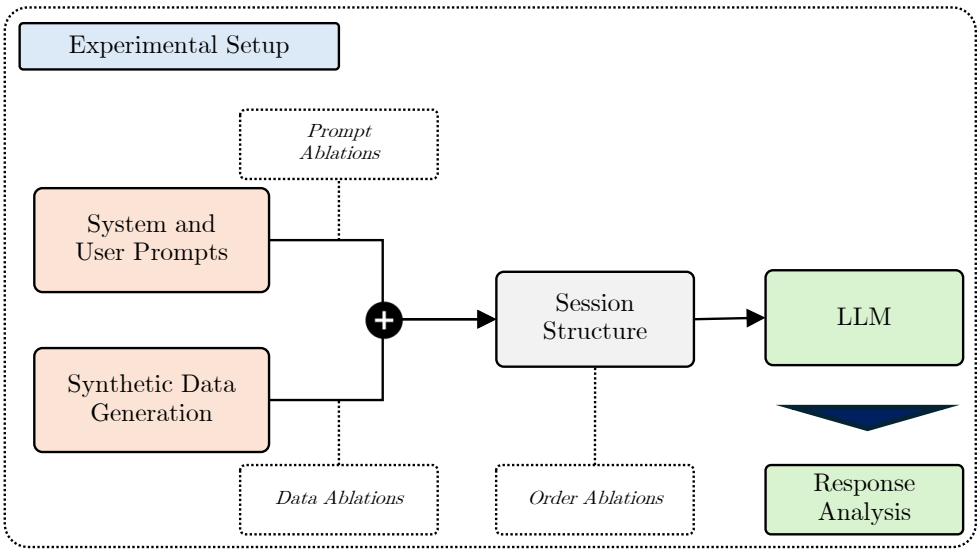

Figure 9: Experimental setup overview

A.2 Interaction with LLMs

A.2.1 Programmatic API

Our interaction with LLMs is through programmatic API calls with temperature fixed at 0.3. As our aim is to probe the natural, emergent behaviour of highly performant LLMs, we instruct models not to use reasoning or chain-of-thought, returning only the final numeric answer with minimal output text. For GPT-5 Mini, reasoning cannot be disabled, so it is set to the lowest reasoning level available. This provides an additional point of comparison, as reasoning-enabled models may behave differently in textual tasks. This is something we see in our experiments involving GPT-5 Mini.

We emphasise the use of API-based LLMs, some of which are closed source, to ensure our pipeline is lightweight and easily extended to new models.

To test modality dependence, we run tasks in text-only, image-only, and text+image conditions. In text-only mode, line-ratio and marker-location tasks are represented using ASCII, while the maze task is described concisely in text. In image-only mode, models receive only the visual stimulus. In multimodal mode, both text and image inputs are given. This allows us to evaluate efficiency in unimodal vs multimodal contexts.

A.2.2 Prompt design

For all tasks, prompts are structured in two parts: a **system prompt** and a **user query**.

- **System prompt**: defines the role of the model (e.g., "You are a line-length ratio estimator."). It specifies the expected output format and instructs the model to *not output reasoning*, but to return only the final numeric estimate (with minimal text if necessary).
- **User query**: provides the stimulus in the chosen modality. In textual mode this is ASCII input (for line ratio and marker tasks) or a concise text description (for maze and subtitle tasks). In image mode only the stimulus image is shown. In multimodal mode both text and image are provided.

A typical prompt for the line-length ratio task (textual mode) is:

```
System prompt:  "You are a line-length ratio
estimator.  Estimate the ratio of the shorter line
to the longer line as a decimal number between 0 and
1.  Do not explain or reason.  Only output the final
answer."
User input:

|-=-=-----                                  |
|-------------------------.------ -------|
```

This design keeps task specification clear and minimises variation in output. For GPT-5 Mini, where reasoning cannot be disabled, we used the lowest reasoning setting. This provides an additional point of comparison, since reasoning-enabled models may behave differently in textual tasks.

For **steering-related ablations**, modifications are made at the system prompt stage. Models may be told that observations are noisy, or given numerical information about the range of past observations. Further details of these manipulations are described in Section A.3.1.

A.3 Ablation Background

Ablation conditions are grouped into three categories: steering-related, noise-related, and context-related. Each modifies the base setup in a controlled way to test specific hypotheses.

A.3.1 Steering-related ablations

Verbal cues

- Modified the system prompt to explicitly tell the model that observations are noisy and that it should act in a Bayesian way.
- Example system prompt:

  > You are a line-length ratio estimator. The given data is noisy and may contain artifacts. You should behave like a Bayesian observer and take into account prior and likelihood in your predictions.

Numerical cues

- Modified the system prompt to provide the numeric range of the past ten observations, encouraging the model to use this information as a prior.
- Example system prompt:

  > You are a line-length ratio estimator. The given data is noisy and may contain artifacts. For 10 previous observations, the values were observed to lie in the range of 0.1 to 0.3.

A.3.2 NOISE-RELATED ABLATIONS

Constant noise: Applied a Gaussian blur to image inputs only, to test whether models adapt estimation behaviour when vision is degraded.

Noise sequence: Introduced gradually increasing Gaussian blur across trials to test whether models downweight visual information as noise grows. Figure 11 shows example input images.

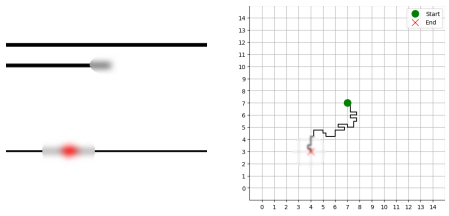

Figure 10: Constant Gaussian noise ablation

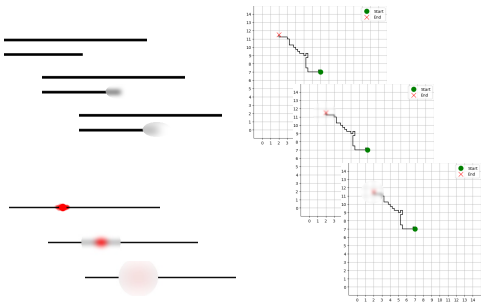

Figure 11: Sequential Gaussian noise ablation

A.3.3 CONTEXT-RELATED ABLATIONS

Shorter Context: Reduced the context window to 3 prior trials, limiting how much past information the model can use.

Longer context Increased the context window to 20 prior trials, maximising available history for the model.

Stimulus order reversal Reversed the order of stimuli to test whether model estimations show strong sequence dependence.

A.4 LLM MODELS STUDIED

Table 3 summarises the key characteristics of the LLMs studied. We chose a diverse set of recent models spanning closed- and open-weight releases, with a range of sizes and architectures. Where possible, we disabled extended-thinking or reasoning controls to probe the models' natural, emergent behaviour. This was feasible for all models except GPT-5 Mini, which only allows adjusting reasoning depth; we set this to the lowest level.

h

Table 3: Comparison of selected LLMs (parameters shown only when vendor/model card publicly discloses them).

Model	Developer	Params	Reasoning controls
Claude 3.7 Sonnet	Anthropic	Undisclosed	Optional "extended thinking"
GPT-5 Mini	OpenAI	Undisclosed	Adjustable depth.
GPT-4o	OpenAI	Undisclosed	N.A.
Llama-4 Maverick	Meta	400B total / 17B active	N.A.
Qwen 2.5 VL 32B	Alibaba	32B	N.A.
Mistral 24B	Mistral	24B	N.A.
Gemini 2.5 Flash Lite	Google DeepMind	Undisclosed	N.A.
Phi 4 Multimodal	Microsoft	Undisclosed	N.A.
Gemma 3 4B	Google DeepMind	4B	N.A.

Notes: We avoid speculative parameter estimates. Public sources: Claude 3.7 Sonnet announcement (Anthropic); GPT-5 Mini (OpenAI docs); Llama-4 Maverick active/total params (Meta); Qwen 2.5-VL 32B model card; Mistral 24B (Mistral docs); Gemini 2.5 Flash-Lite (Google); Phi-4 Multimodal (Microsoft HF card); Gemma 3 model card.

A.5 Human feedback collection

We collected data from human subjects on our main tasks to establish a calibration benchmark. The questions are hosted on a web platform, and users can complete them with their phone or computer.

Only two ablations were used for human feedback collection: constant noise and longer context.

Figure 12 shows two screenshots of the web platform.

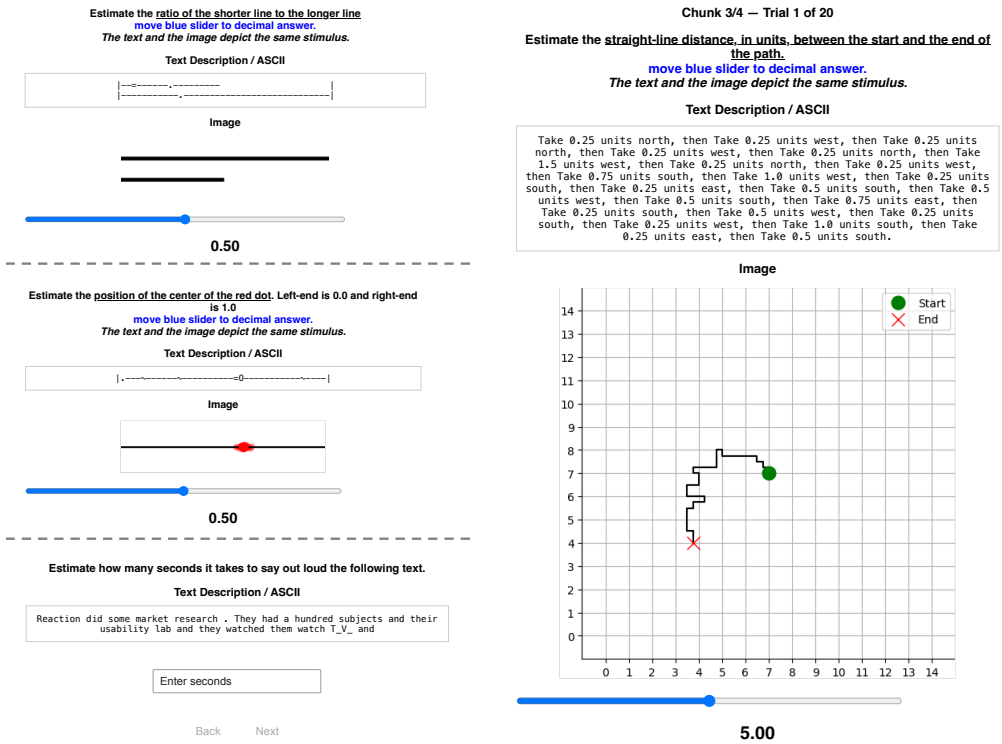

Figure 12: Human feedback collection website screenshot

A.6 BAYES CUE COMBINATION MODELS

Under Bayesian assumptions, the optimal linear combination of two noisy modality estimates is obtained by weighting them according to their relative reliabilities (inverse variances. See also Ernst & Banks (2002)). We consider two versions. Non-oracle and oracle models. They differ in whether the cue combination is modelled with or without access to ground truth.

- **Non-oracle:** The model combines the two modality estimates ($\mu^{(1)}$ and $\mu^{(2)}$) by inverse-variance weighting,
$$\mu = \frac{\tau_1}{\tau_1 + \tau_2}\mu^{(1)} + \frac{\tau_2}{\tau_1 + \tau_2}\mu^{(2)},$$
where $\tau_i = 1/\sigma_i^2$ are the precisions of estimates from the corresponding modality. Crucially, the model does not assume access to ground truth. It only uses the variance of each modality estimate to compute the above weighting.
- **Oracle:** In this case, we first calibrate the modality-specific estimates ($\mu^{(1)}$ and $\mu^{(2)}$) by fitting gain and offset parameters to the ground truth for each modality. After calibration, the estimates ($\mu'^{(1)}$ and $\mu'^{(2)}$) are combined using the generalised least squares solution based on the residual loss covariance (Σ):
$$\mu = \frac{\mathbf{1}^\top \Sigma^{-1} \boldsymbol{\mu'}}{\mathbf{1}^\top \Sigma^{-1} \mathbf{1}},$$
where $\boldsymbol{\mu'} = \begin{bmatrix} \mu'^{(1)} \\ \mu'^{(2)} \end{bmatrix}$ and Σ is the 2×2 covariance matrix of the modality estimates. This accounts for both differing reliabilities and cross-modal correlations, yielding the optimal linear unbiased estimator given access to the true values.

Although these models specify optimal *linear* integration strategies, it is important to note that LLMs may, in principle, outperform these baselines if they achieve more flexible, nonlinear forms of cue integration. Such nonlinear integration is possible given the architecture of modern LLMs.

A.7 FACTOR ANALYSIS DETAILS

We fit many behavioural model variants that differ along interpretable *factors* (e.g., BAYESIAN vs NON-BAYESIAN; WEBER vs NON-WEBER; SEQUENTIAL update). Because these variants partly overlap in purpose, naively summing or averaging likelihoods would (i) reward families that contain more variants, or (ii) dilute good variants by pooling with weak ones. We therefore compare *factors* while treating all other dimensions as *nuisance*.

Procedure Let $f \in \{\text{BAYESIAN}, \text{WEBER}, \text{SEQUENTIAL}\}$ be the factor of interest, and let $\mathcal{N}(f)$ denote the set of nuisance factors for this comparison (chosen to be agnostic to f; see example below).

1. **Transform AIC to likelihood.** For each fitted variant m, first compute $\Delta\text{AIC}(m)$ (defined as the difference between m's AIC and the minimum AIC among all variants) and then compute the transformed quantity below:
$$L(m) \;\propto\; \exp\big(-\tfrac{1}{2}\,\Delta\text{AIC}(m)\big),$$
2. **Group by nuisance "cells".** Group behavioural models by every combination of values in $\mathcal{N}(f)$. Each group is a cell c.
3. **Best-in-cell for each level of f.** Within each cell c, take the *maximum* likelihood among variants where $f = \texttt{True}$ and among variants where $f = \texttt{False}$:
$$L^{(c)}_{\text{True}} = \max_{m \in c,\ f(m)=\texttt{True}} L(m), \qquad L^{(c)}_{\text{False}} = \max_{m \in c,\ f(m)=\texttt{False}} L(m).$$
Using the max avoids penalising a family for having many weak sub-variants.

4. **Equal-weight across cells.** For fairness, average *equally* across cells where both levels are present (intersection):

$$\bar{L}_{\text{True}} = \frac{1}{|C|} \sum_{c \in C} L^{(c)}_{\text{True}}, \qquad \bar{L}_{\text{False}} = \frac{1}{|C|} \sum_{c \in C} L^{(c)}_{\text{False}},$$

where $C = \{c : \ L^{(c)}_{\text{True}}, L^{(c)}_{\text{False}}$ both defined above in step 3.$\}$.

5. **Compute evidence.** Report the factor-level probability

$$P(f{=}\texttt{True} \mid \text{data}) \ = \ \frac{\bar{L}_{\text{True}}}{\bar{L}_{\text{True}} + \bar{L}_{\text{False}}},$$

and similarly for `False`.

In this report when we refer to ***factor evidence***, we are always referring to evidence computed from this procedure.

Example: BAYESIAN vs NON-BAYESIAN. For f = BAYESIAN we take $\mathcal{N}(f) = \{\text{WEBER}\}$ only. The SEQUENTIAL and GAIN variants exist exclusively within the Bayesian family; conditioning on them would create empty cells on the non-Bayesian side. Thus, within each WEBER cell we compare the best Bayesian variant (possibly sequential/gain/log) against the best non-Bayesian variant, average equally over cells, and form the head-to-head probability. Below table shows the procedure schematically.

	WEBER cell	
	`False`	`True`
Best Bayesian in cell	$L^{(c)}_{\text{True}}$	$L^{(c)}_{\text{True}}$
Best non-Bayesian in cell	$L^{(c)}_{\text{False}}$	$L^{(c)}_{\text{False}}$

Average equally across cells, then compute $P = \bar{L}_{\text{True}}/(\bar{L}_{\text{True}} + \bar{L}_{\text{False}})$.

Notes on fairness and robustness. (i) Equal cell weighting prevents families with many variants from accruing more probability mass simply by proliferation. (ii) Using the intersection of cells avoids bias from missing combinations.

A.8 BCS FITTING DETAILS

We fit the static Bayesian observer model in all cases and with data from modalities according to the below:

- **Noise:** evaluate w_{prior} from the *image-only* modality, since noise is injected only into the image channel and multimodal fits would confound reweighting of text input.
- **Steering and Context:** evaluate w_{prior} from the *multimodal* fit, as these manipulations affect both modalities.

A.9 MODEL PERFORMANCE AND BAYESIAN FACTOR EVIDENCE

Figures 13, 14, 15 and 16 show the NRMSE performance and Bayesian factor evidence for all models across all tasks and modalities. For the multimodal tasks, in their corresponding figures, metrics by modality is shown over the three rows.

Notice that not all models perform better in multimodal conditions than in unimodal conditions (Llama-4 Maverick is the outlier, it achieves its best NRMSE in multimodal mode on all tasks).

A.10 GPT-5 MINI CUE COMBINATION MODEL FITS

GPT-5 Mini's cue-combination performance is poor despite its very strong NRMSE performance. Figure 15 shows the NRMSE performance for each model in all three modalities for the maze distance estimation task. We see that GPT-5 Mini's unimodal text performance is nearly perfect

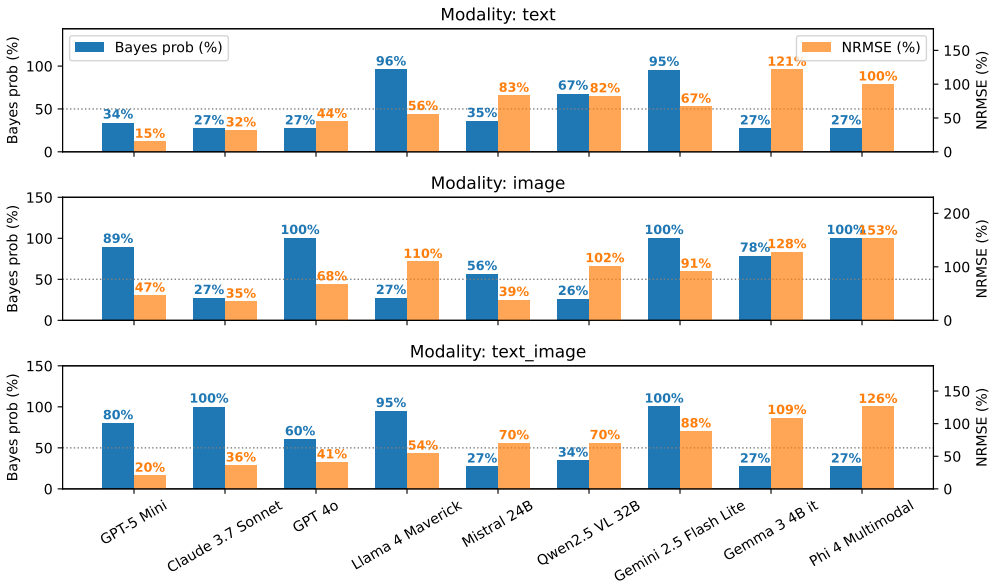

Figure 13: Line length ratio estimation task. NRMSE and Bayes factor evidence for unimodal text, unimodal image and multimodal inputs.

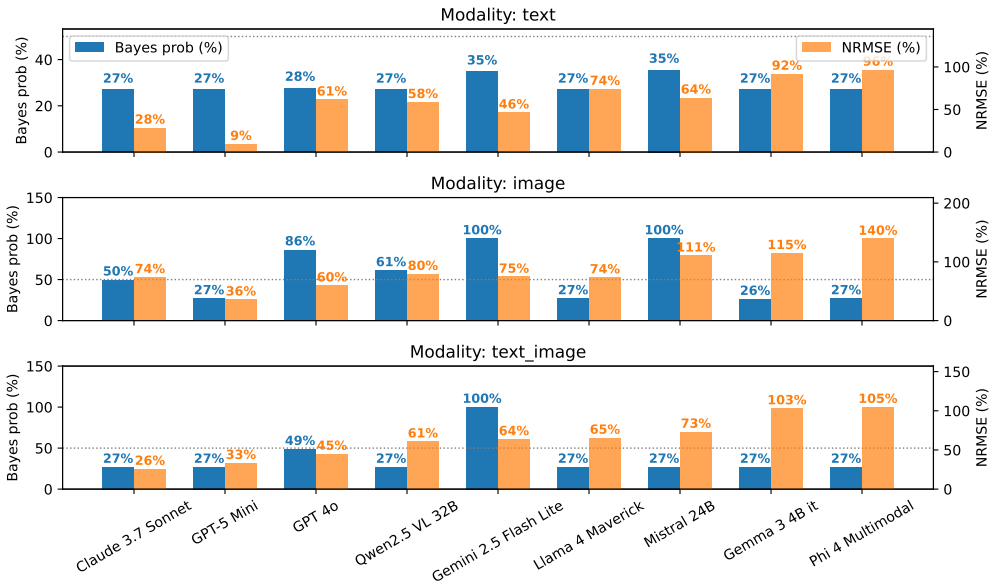

Figure 14: Marker location estimation task. NRMSE and Bayes factor evidence for unimodal text, unimodal image and multimodal inputs.

(at 0.01 NRMSE), while its unimodal image performance is much worse (at 0.2 NRMSE, despite already being the best across models). Because of this, the Bayes-optimal linear combination would imply a nearly zero weighing on the image input. However, the multimodal performance does not follow this trend, indicating that the model prediction is still affected by the image input.

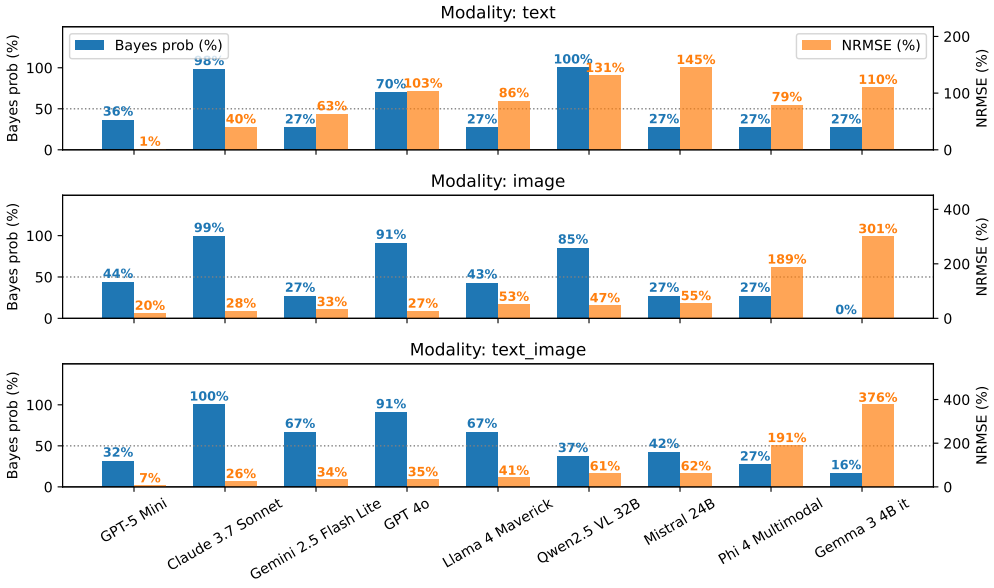

Figure 15: Maze distance estimation task. NRMSE and Bayes factor evidence for unimodal text, unimodal image and multimodal inputs.

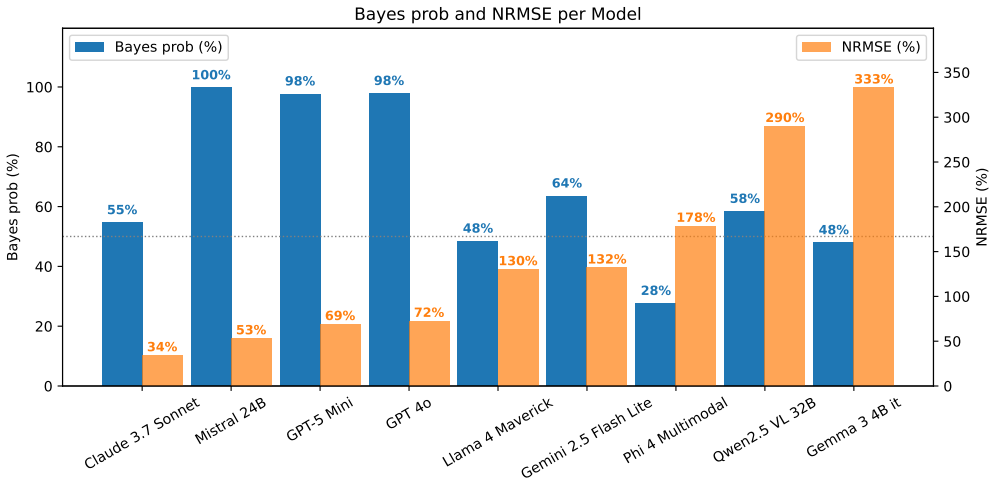

Figure 16: Subtitle duration estimation task. NRMSE and Bayes factor evidence.

A.11 Further model variants

Studies such as (Nieder & Miller, 2003; Nover et al., 2005) found in human studies that the human brain encodes many different magnitudes using a logarithmic scale. To test if this phenomena apply in LLM, we explored variants of models where a logarithmic transform is applied to the stimulus values.

For the Bayesian model we also added variants with an affine transform after the estimate is computed, to account for any potential gain biases. This is not needed for the linear models as it is captured by the gradient parameter.

Note that for all these variants, the additional parameters will penalise AIC and therefore help guard against artificial model evidence inflation by more complex models.

A.11.1 LOGARITHMIC TRANSFORM

In some model variants, a logarithmic transform is applied to the stimulus or response space before fitting our behavioural models above. This is motivated by standard assumptions in psychophysics that humans internally represent magnitudes on a log scale.

Thus the transformed stimulus x'_t from the raw input x_t is

$$x'_t = \log(x_t + \epsilon),$$

with a small ϵ ensuring numerical stability. Log-transform variants are considered for both the linear and Bayesian observer models.

A.11.2 AFFINE TRANSFORM

For Bayesian models, we additionally allow affine deviations of the posterior estimate, corresponding to a gain factor $g \in \mathbb{R}^+$ and an additive offset $\delta \in \mathbb{R}$. The raw posterior mean μ_t from the model estimate is transformed to $\tilde{\mu}_t$ as

$$\tilde{\mu}_t = g\,\mu_t + \delta.$$

The LLM response y_t in these variants is generated as below, where σ^2_{dec} is again a free parameter fitted during the model fitting stage:

$$y_t \sim \mathcal{N}(\tilde{\mu}_t,\, \sigma^2_{\text{dec}}).$$

This captures systematic deviations from the normative Bayesian solution, such as under- or over-weighting of evidence and constant response bias. Note that for linear models this is not required as it is already captured by the slope and offset parameters.