Bayesian Evals

December 2024

Evaluating AI Systems is hard

- Training set contamination

See Anthropic blog: “Challenges in evaluating AI systems” (2023)

Evaluating AI Systems is hard

- Training set contamination
- Differences in formatting/prompting lead to changes in performance

See Anthropic blog: “Challenges in evaluating AI systems” (2023)

Evaluating AI Systems is hard

- Training set contamination
- Differences in formatting/prompting lead to changes in performance
- *<u>So many eval datasets to use, so little time– e.g. big bench has 204 tasks</u>*

See Anthropic blog: “Challenges in evaluating AI systems” (2023)

Evaluating AI Systems is hard

- Training set contamination
- Differences in formatting/prompting lead to changes in performance
- *So many eval datasets to use, so little time– e.g. big bench has 204 tasks*
- Can try model-generated evals but then you get an ouroboros of evals

See Anthropic blog: “Challenges in evaluating AI systems” (2023)

Evaluating AI Systems is hard

- Training set contamination
- Differences in formatting/prompting lead to changes in performance
- *So many eval datasets to use, so little time– e.g. big bench has 204 tasks*
- Can try model-generated evals but then you get an ouroboros of evals
- *How best to model uncertainty in evals?*

See Anthropic blog: “Challenges in evaluating AI systems” (2023)

Evaluating AI Systems is hard

- Training set contamination
- Differences in formatting/prompting lead to changes in performance
- *So many eval datasets to use, so little time– e.g. big bench has 204 tasks*
- Can try model-generated evals but then you get an ouroboros of evals
- *How best to model uncertainty in evals?*
- How to test for subjective quality of generations (not MCQA)

See Anthropic blog: “Challenges in evaluating AI systems” (2023)

Evaluating AI Systems is hard

- Training set contamination
- Differences in formatting/prompting lead to changes in performance
- *So many eval datasets to use, so little time– e.g. big bench has 204 tasks*
- Can try model-generated evals but then you get an ouroboros of evals
- *How best to model uncertainty in evals?*
- How to test for subjective quality of generations (not MCQA)
    - A/B tests (e.g. chatbot arena -> Elo scores) – expensive, difficult to model uncertainty
    - Red teaming – difficult, not standardised

See Anthropic blog: "Challenges in evaluating AI systems" (2023)

OLMES (Open Language Model Evaluation Standard) - Gu et al.

- Considers the uncertainty between different experiment setups

- But not the uncertainty inside each eval experiment

- (that is, inter-model uncertainty but not intra-model uncertainty)

Table 1: Scores reported in different references for LLM performances on ARC-CHALLENGE and OPENBOOKQA. Scores indicated with † are using multiple-choice formulation (MCF) rather than "cloze" formulation (CF) (see Section 2.1 for definitions). Entries with "?" denote either undocumented or mixed approaches across models. Different references use different evaluation setups, some of which are not fully specified, so conclusions about which models perform best are not reproducible.

	ARC-CHALLENGE Evaluations:							OPENBOOKQA Evaluations:					
Model↓	**Ref1**	**Ref2**	**Ref3**	**Ref4**	**Ref5**	**Ref6**	**OLMES**	**Ref2**	**Ref4**	**Ref5**	**Ref7**	**Ref8**	**OLMES**
MPT-7B	47.7	42.6			46.5		45.7	51.4		48.6			52.4
RPJ-INCITE-7B	46.3				42.8		45.3			49.4			49.0
Falcon-7B	47.9	42.4		44.5	47.5		49.7	51.6	44.6	53.0		26.0†	55.2
Mistral-7B	60.0		55.5	54.9			78.6†				52.2	77.6†	80.6†
Llama2-7B	53.1	45.9	43.2	45.9	48.5	53.7†	54.2	58.6	58.6	48.4	58.6	54.4†	57.8
Llama2-13B	59.4	49.4	48.8	49.4		67.6†	67.3†	57.0	57.0		57.0	63.4†	65.4†
Llama3-8B	60.2					78.6†	79.3†					76.6†	77.2†
Num shots	25	0	0	0	0	25	5	0	0	0	0	5	5
Curated shots	No					No	Yes					No	Yes
Formulation	RC	RC	RC?	RC	RC	MC	MC/RC	RC	RC	RC	RC	MC	MC/RC
Normalization	char	char	?	char?	pmi	none	none/pmi	pmi	pmi?	pmi	pmi?	none	none/pmi

Ref	**Reference citation**	**Ref**	**Reference citation**
Ref1	HF Open LLM Leaderboard (Beeching et al., 2023)	**Ref5**	OLMo paper (Groeneveld et al., 2024)
Ref2	Llama2 paper (Touvron et al., 2023a)	**Ref6**	Llama3 model card (AI@Meta, 2024)
Ref3	Mistral 7B (Jiang et al., 2023)	**Ref7**	Gemma paper (Gemma Team et al., 2024)
Ref4	Falcon paper (Almazrouei et al., 2023)	**Ref8**	HELM Lite Leaderboard (Liang et al., 2023)

BetterBench – Reuel et al.

- Methodology for evaluating quality of benchmarks

- “14 out of 24 benchmarks did not perform multiple evaluations of the same model or report statistical significance or uncertainty of results”

- “an average score of 5.62 [out of 15] on *Reporting statistical significance*”

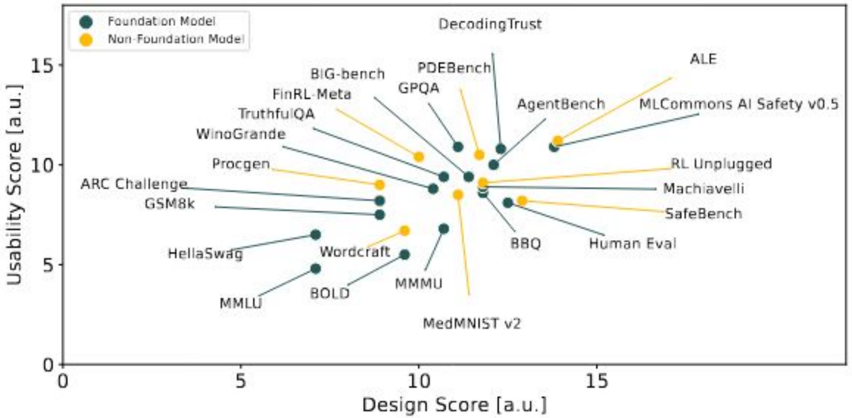
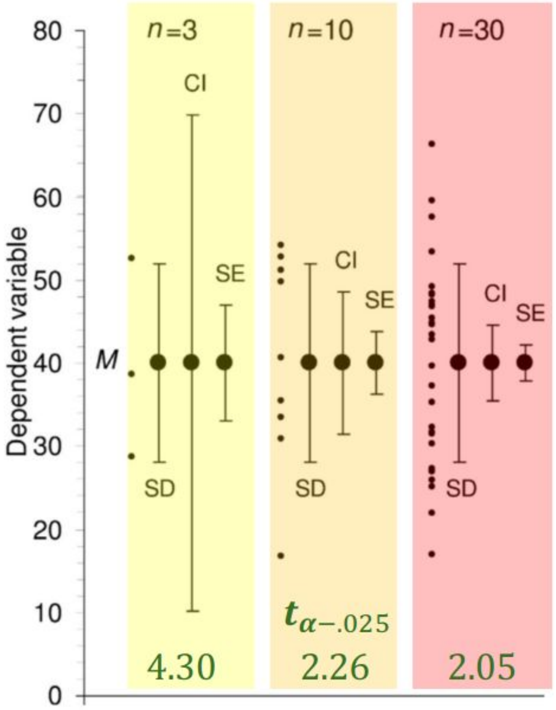


Figure 7: Design and usability score for all 24 assessed benchmarks. The usability score is the weighted average of the implementation, documentation, and maintenance scores. Benchmarks were split into foundation model and non-foundation model benchmarks, depending on the model group they’re targeting.

What to report in your evals

- Accuracy (overall and per-task) averaged over n runs

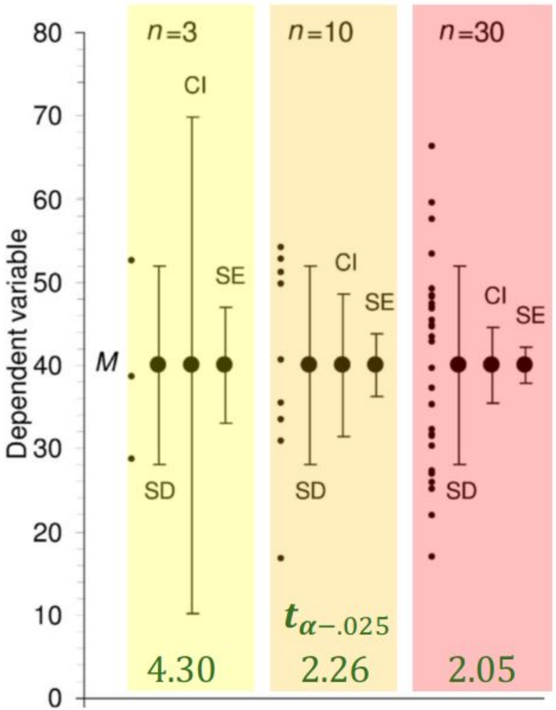


Cumming, Fidler, & Vaux (2007)

See Neurips slides by Hermann et al. (2024) for more

What to report in your evals

- Accuracy (overall and per-task) averaged over n runs
- With error bars:
  - Using standard error rather than standard deviation:
    - Uncertainty in the estimate of the mean, not just the spread of data

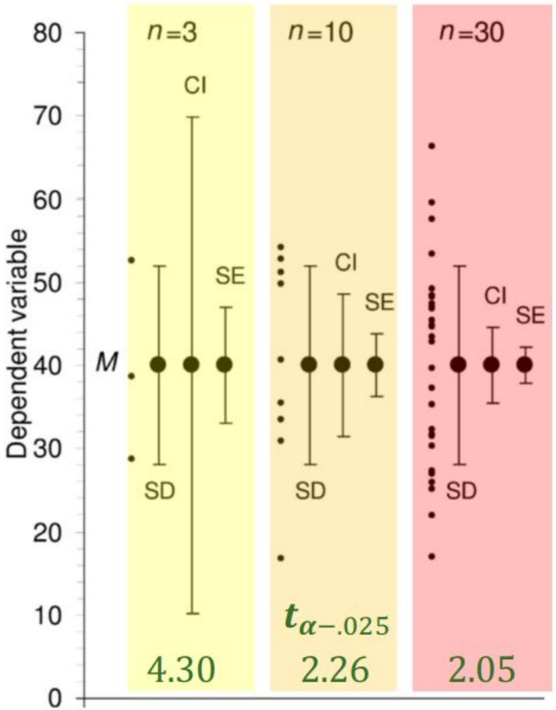


Cumming, Fidler, & Vaux (2007)

See Neurips slides by Hermann et al. (2024) for more

What to report in your evals

- Accuracy (overall and per-task) averaged over n runs
- With error bars:
    - Using standard error rather than standard deviation:
        - Uncertainty in the estimate of the mean, not just the spread of data
- How can we represent the facts that
    - Each benchmark/task might be of different quality?
    - Each model may be more/less capable/consistent on a different subset of benchmarks/tasks?

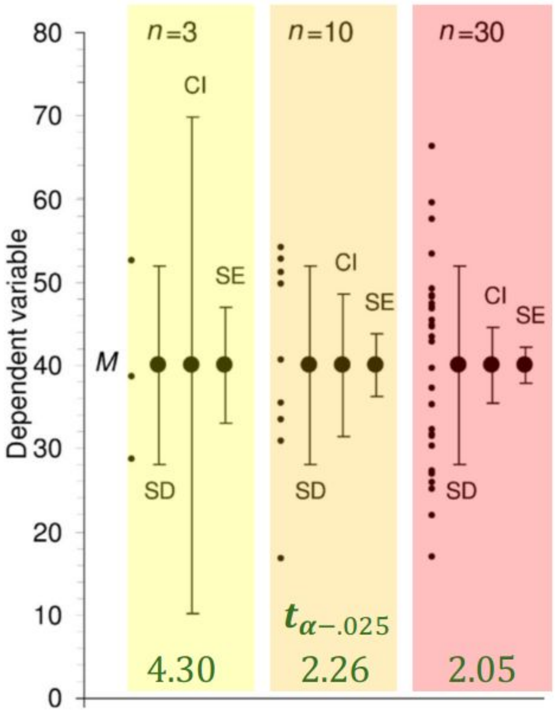


Cumming, Fidler, & Vaux (2007)

See Neurips slides by Hermann et al. (2024) for more

What to report in your evals

- Accuracy (overall and per-task) averaged over n runs
- With error bars:
    - Using standard error rather than standard deviation:
        - Uncertainty in the estimate of the mean, not just the spread of data
- How can we represent the facts that
    - Each benchmark/task might be of different quality?
    - Each model may be more/less capable/consistent on a different subset of benchmarks/tasks?
- How can we use these error bars to consider the probability with which we can say “Model A > Model B”?



Cumming, Fidler, & Vaux (2007)

See Neurips slides by Hermann et al. (2024) for more

Set up

Suppose we have n questions, and our model achieves scores $\{s_i\}_{i=1}^n$ (between 0 and 1) where s_i can be decomposed into a mean component x_i and a zero-mean random component ϵ_i.

$$s_i = x_i + \epsilon_i$$

We want to infer $\mu = \mathbb{E}[s] = \mathbb{E}[x]$

Frequentist Error Bars

Let $\bar{s} = \frac{1}{n} \sum_i s_i$. Then by the law of large numbers: $\hat{\mu} = \bar{s}$

Frequentist Error Bars

Let $\bar{s} = \frac{1}{n}\sum_i s_i$. Then by the law of large numbers: $\hat{\mu} = \bar{s}$

CLT: $\mathrm{SE}_{\mathrm{CLT}} = \sqrt{\mathrm{Var}(s)/n} = \sqrt{\left(\frac{1}{n-1}\sum_i (s_i - \bar{s})^2\right)/n}$

Frequentist Error Bars

Let $\bar{s} = \frac{1}{n}\sum_i s_i$. Then by the law of large numbers: $\hat{\mu} = \bar{s}$

CLT: $\mathrm{SE}_{\mathrm{CLT}} = \sqrt{\mathrm{Var}(s)/n} = \sqrt{\left(\frac{1}{n-1}\sum_i (s_i - \bar{s})^2\right)/n}$

If $s_i \in \{0, 1\}$ then this becomes $\mathrm{SE}_{\mathrm{Bernoulli}} = \sqrt{\hat{s}(1-\hat{s})/n}$

Frequentist Error Bars

Let $\bar{s} = \frac{1}{n}\sum_i s_i$. Then by the law of large numbers: $\hat{\mu} = \bar{s}$

CLT: $\mathrm{SE}_{\mathrm{CLT}} = \sqrt{\mathrm{Var}(s)/n} = \sqrt{\left(\frac{1}{n-1}\sum_i (s_i - \bar{s})^2\right)/n}$

If $s_i \in \{0, 1\}$ then this becomes $\mathrm{SE}_{\mathrm{Bernoulli}} = \sqrt{\hat{s}(1-\hat{s})/n}$

Obtain 95% confidence intervals as $\mathrm{CI}_{95\%} = \bar{s} \pm 1.96 \times \mathrm{SE}_{\mathrm{CLT}}$

(if you do the data collection 100 times, 95 of those times you'll get $\mu \in \mathrm{CI}_{95\%}$)

Anthropic Paper “Adding Error Bars to Evals” (Miller, 2024)

Points out that the LLama 3 (Dubey et al., 2024) paper reports only $\mathrm{SE}_{\mathrm{Bernoulli}}$, even when s_i can take on values in $(0, 1)$ (e.g. F1 score). This leads to confidence intervals that are too wide.


Meta

The Llama 3 Herd of Models

Llama Team, AI @ Meta[1]

Anthropic Paper “Adding Error Bars to Evals” (Miller, 2024)

Points out that the LLama 3 (Dubey et al., 2024) paper reports only $\mathrm{SE}_{\mathrm{Bernoulli}}$, even when s_i can take on values in $(0,1)$ (e.g. F1 score). This leads to confidence intervals that are too wide.

Meta

The Llama 3 Herd of Models

Llama Team, AI @ Meta[1]

Significance estimates. Benchmark scores are estimates of a model's true performance. These estimates have variance because benchmark sets are finite samples drawn from some underlying distribution. We follow Madaan et al. (2024b) and report on this variance via 95% confidence intervals (CIs), assuming that benchmark scores are Gaussian distributed. While this assumption is incorrect (*e.g.*, benchmark scores are bounded), preliminary bootstrap experiments suggest CIs (for discrete metrics) are a good approximation:

$$CI(S) = 1.96 \times \sqrt{\frac{S \times (1 - S)}{N}}.$$

Herein, S is the observed benchmark score (*e.g.*, accuracy or EM) and N the sample size of the benchmark. We omit CIs for benchmark scores that are not simple averages. We note that because subsampling is not the only source of variation, our CI values lower bound the actual variation in the capability estimate.

Anthropic Paper “Adding Error Bars to Evals” (Miller, 2024)

In reality, questions come in groups from different tasks (e.g. boolean_expression, date_understanding from BBH), so the IID assumption of the CLT is violated.

Anthropic Paper "Adding Error Bars to Evals" (Miller, 2024)

In reality, questions come in groups from different tasks (e.g. boolean_expression, date_understanding from BBH), so the IID assumption of the CLT is violated.

Therefore suggests using clustered standard error, where $s_{i,c}$ is the ith question from cluster/task c

$$\mathrm{SE}_{\text{clustered}} = \left(\mathrm{SE}^2_{\text{C.L.T.}} + \frac{1}{n^2} \sum_{c} \sum_{i} \sum_{j \neq i} (s_{i,c} - \bar{s})(s_{j,c} - \bar{s}) \right)^{1/2}$$

Anthropic Paper “Adding Error Bars to Evals” (Miller, 2024)

In reality, questions come in groups from different tasks (e.g. boolean_expression, date_understanding from BBH), so the IID assumption of the CLT is violated.

Therefore suggests using clustered standard error, where $s_{i,c}$ is the ith question from cluster/task c

$$\mathrm{SE}_{\text{clustered}} = \left(\mathrm{SE}^2_{\text{C.L.T.}} + \frac{1}{n^2} \sum_c \sum_i \sum_{j \neq i} (s_{i,c} - \bar{s})(s_{j,c} - \bar{s}) \right)^{1/2}$$

	$\mathrm{SE}_{\text{clustered}}$	$\mathrm{SE}_{\text{C.L.T.}}$	Ratio
DROP	(1.34)	(0.44)	3.05
RACE-H	(0.51%)	(0.46%)	1.10
MGSM	(1.62%)	(0.86%)	1.88

Table 4: Clustered and naive standard errors computed on two popular evals using Anthropic models (non-fictional numbers). Analyzing the same data, clustered standard errors can be over 3X larger than naive standard errors.

“sliding scale between cases where scores within a cluster are perfectly correlated ([...] each cluster acts as a single indep. observation) and perfectly uncorrelated”

Anthropic Paper “Adding Error Bars to Evals” (Miller, 2024)

How to compare two models, A and B?

Naively: $\hat{\mu}_{A-B} = \hat{\mu}_A - \hat{\mu}_B$ $\quad \mathrm{SE}_{A-B} = \sqrt{\mathrm{SE}_A^2 + \mathrm{SE}_B^2}$ $\quad \mathrm{CI}_{A-B,95\%} = \hat{\mu}_{A-B} \pm 1.96 \times \mathrm{SE}_{A-B}$

Anthropic Paper "Adding Error Bars to Evals" (Miller, 2024)

How to compare two models, A and B?

Naively: $\hat{\mu}_{A-B} = \hat{\mu}_A - \hat{\mu}_B$ $\quad \mathrm{SE}_{A-B} = \sqrt{\mathrm{SE}_A^2 + \mathrm{SE}_B^2}$ $\quad \mathrm{CI}_{A-B,95\%} = \hat{\mu}_{A-B} \pm 1.96 \times \mathrm{SE}_{A-B}$

But if you know the set of questions used for A and B's evals you can do paired analysis: $s_{A-B,i} = s_{A,i} - s_{B,i}$ $\quad \bar{s}_{A-B} = \bar{s}_A - \bar{s}_B$

$$\mathrm{SE}_{A-B,\text{paired}} = \sqrt{\mathrm{Var}(s_{A-B})/n} = \sqrt{\left(\frac{1}{n-1}\sum_i (s_{A-B,i} - \bar{s}_{A-B})^2\right)/n}$$

Anthropic Paper “Adding Error Bars to Evals” (Miller, 2024)

How to compare two models, A and B?

Naively: $\hat{\mu}_{A-B} = \hat{\mu}_A - \hat{\mu}_B \quad \mathrm{SE}_{A-B} = \sqrt{\mathrm{SE}_A^2 + \mathrm{SE}_B^2} \quad \mathrm{CI}_{A-B,95\%} = \hat{\mu}_{A-B} \pm 1.96 \times \mathrm{SE}_{A-B}$

But if you know the set of questions used for A and B’s evals you can do paired analysis: $s_{A-B,i} = s_{A,i} - s_{B,i} \qquad \bar{s}_{A-B} = \bar{s}_A - \bar{s}_B$

$$\mathrm{SE}_{A-B,\text{paired}} = \sqrt{\mathrm{Var}(s_{A-B})/n} = \sqrt{\left(\frac{1}{n-1}\sum_i (s_{A-B,i} - \bar{s}_{A-B})^2\right)/n}$$

Or with clustering:

$$\mathrm{SE}_{A-B,\text{paired,clustered}} = \frac{1}{n}\left(\sum_c \sum_i \sum_j (s_{A-B,i,c} - \bar{s}_{A-B})(s_{A-B,j,c} - \bar{s}_{A-B})\right)^{1/2}$$

How are we modelling the process of evals?

So far just used CLT/Gaussianity assumption – can we do better?

Anthropic paper suggests (briefly) $x_i \sim \mathcal{U}[0,1]$ and $s_i \sim Ber(x_i)$ so $\epsilon_i = 1 - x_i$ with probability x_i and $\epsilon_i = -x_i$ with probability $1 - x_i$.

How are we modelling the process of evals?

So far just used CLT/Gaussianity assumption – can we do better?

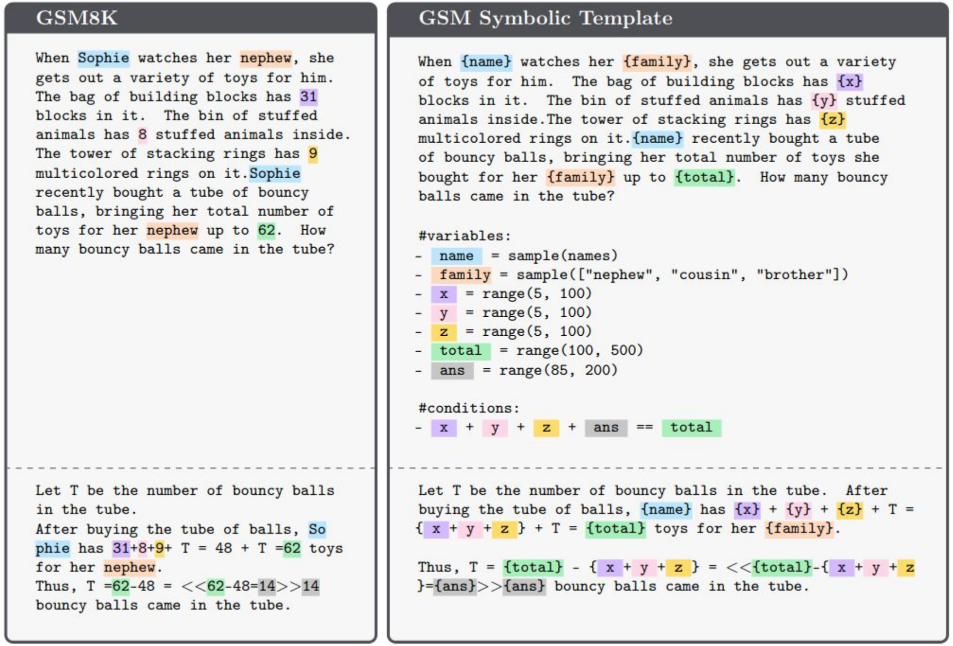
Anthropic paper suggests (briefly) $x_i \sim \mathcal{U}[0, 1]$ and $s_i \sim Ber(x_i)$ so $\epsilon_i = 1 - x_i$ with probability x_i and $\epsilon_i = -x_i$ with probability $1 - x_i$.

Desi’s blogs go into more detail with Binomial modelling.

Desi’s Blog

Examines “GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models” by Mirzadeh et al. (2024) which:

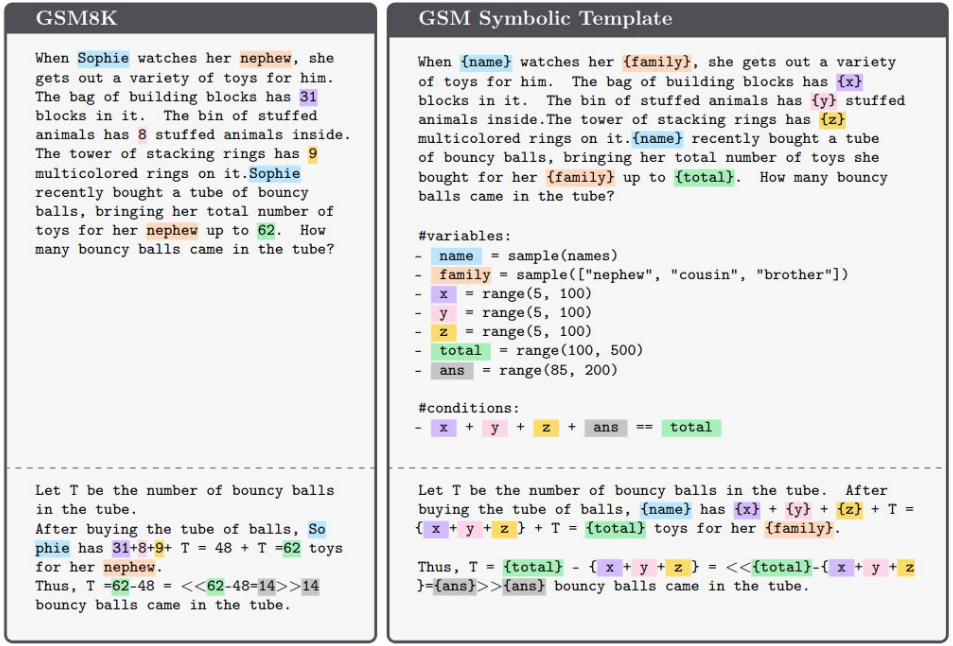
- Takes the popular maths benchmark GSM8K and creates a new version GSM-Symbolic in which we can have “same questions, different numbers”



Desi’s Blog

Examines “GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models” by Mirzadeh et al. (2024) which:

- Takes the popular maths benchmark GSM8K and creates a new version GSM-Symbolic in which we can have “same questions, different numbers”
- Claims that:
    1. Models exhibit worse performance on GSM-Symbolic than GSM8K; and
    2. This implies that “current LLMs are not capable of genuine logical reasoning; instead, they attempt to replicate the reasoning steps observed in their training data”



But claim 1 needs backing up

- Mirzadeh et al. create 50 versions of GSM-Symbolic and run evals for various models
- Compare against a point estimate for GSM8K
- Assumption: if a model is doing ‘logic’ then its performance on these datasets should be equivalent.

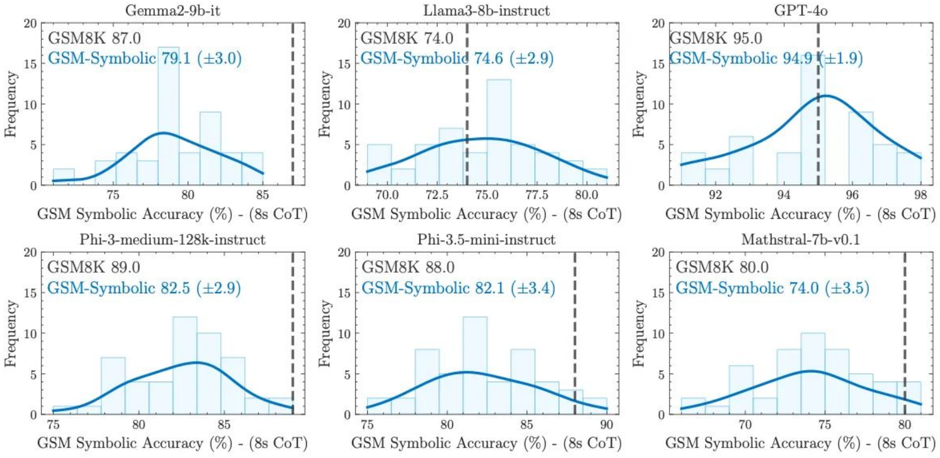


Figure 2: The distribution of 8-shot Chain-of-Thought (CoT) performance across 50 sets generated from `GSM-Symbolic` templates shows significant variability in accuracy among all state-of-the-art models. Furthermore, for most models, the average performance on `GSM-Symbolic` is lower than on `GSM8K` (indicated by the dashed line). Interestingly, the performance of `GSM8K` falls on the right side of the distribution, which, statistically speaking, should have a very low likelihood, given that `GSM8K` is basically a single draw from `GSM-Symbolic`.

But claim 1 needs backing up

Assume each model m has a probability of success p_m (the same across all N questions).

Then the number of correct answers is $\mathrm{Bin}(N, p_m)$.

So we can get (Wilson) confidence intervals.

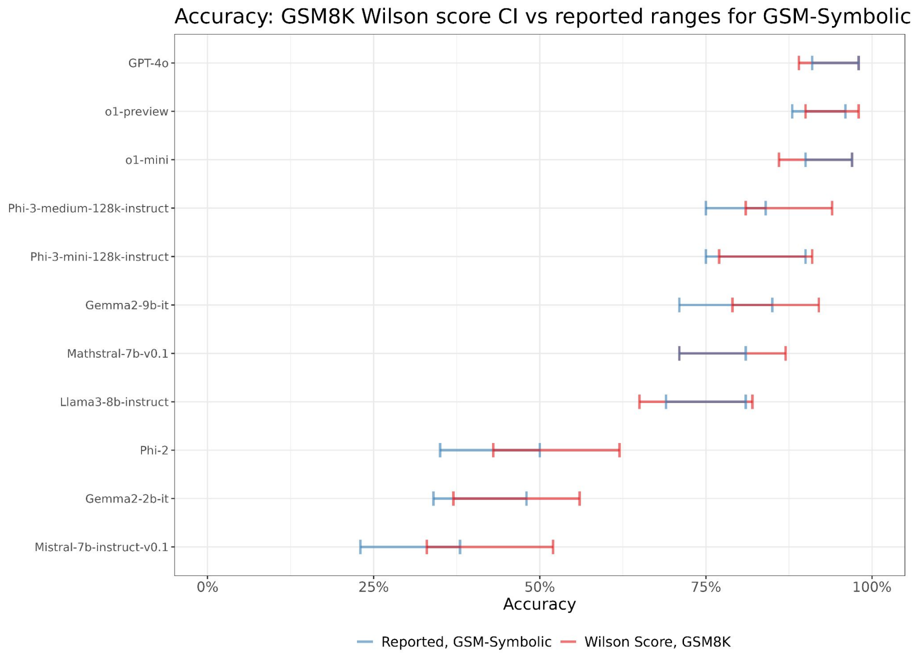
Now it's less obvious that model performance on GSM-Symbolic *is* actually worse than on GSM8K.

n_s= #success

n_f= #fail

$$z_\alpha \approx \frac{(p - \hat{p})}{\sigma_n}$$

$$p \underset{\approx\alpha}{\in} \frac{n_s + \frac{1}{2} z_\alpha^2}{n + z_\alpha^2} \pm \frac{z_\alpha}{n + z_\alpha^2} \sqrt{\frac{n_s n_f}{n} + \frac{z_\alpha^2}{4}}$$



Is the drop in performance statistically significant?

Let p_m^{8K} be the true probability of success for model m on GSM8K and p_m^{Symb} be the true probability of success on GSM-Symbolic.

Is the drop in performance statistically significant?

Let p_m^{8K} be the true probability of success for model m on GSM8K and p_m^{Symb} be the true probability of success on GSM-Symbolic.

Run Fisher's exact test on two-sided and one-sided hypotheses:

$$H_0 : p_m^{8K} = p_m^{\text{Symb}}$$

$$H_1^{\text{two-sided}} : p_m^{8K} \neq p_m^{\text{Symb}}$$

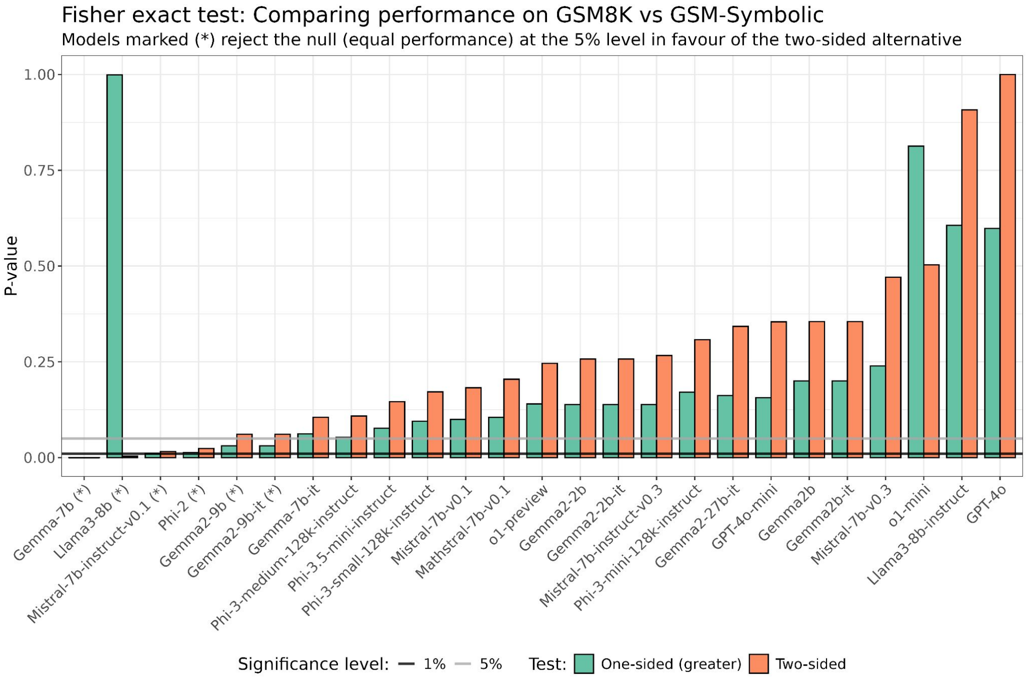
$$H_0 : p_m^{8K} = p_m^{\text{Symb}}$$

$$H_1^{\text{one-sided}} : p_m^{8K} > p_m^{\text{Symb}}$$

Is the drop in performance statistically significant? … Maybe

It turns out there is some (weak) evidence that the models *are* performing worse on GSM-Symbolic compared to GSM8K

(especially when you include irrelevant information in the questions)



Bayesian Error Bars

- Can we do similar eval analysis/modelling with Bayes?
    - Yes!

Bayesian Error Bars

- Can we do similar eval analysis/modelling with Bayes?
  - Yes!
- Benefits:
  - Rather than confidence intervals (95% of which may contain the true parameter) we can use credible intervals and make direct statements such as “the true parameter lies in here with 95% probability”.
  - Allow for complex hierarchical modelling
  - Doesn’t rely on CLT (IID and large N assumptions)
  - Avoid complicated construction of frequentist setups (p-values shenanigans)

Bayesian Error Bars

- Can we do similar eval analysis/modelling with Bayes?
  - Yes!
- Benefits:
  - Rather than confidence intervals (95% of which may contain the true parameter) we can use credible intervals and make direct statements such as “the true parameter lies in here with 95% probability”.
  - Allow for complex hierarchical modelling
  - Doesn’t rely on CLT (IID and large N assumptions)
  - Avoid complicated construction of frequentist setups (p-values shenanigans)
- Drawbacks:
  - Have to choose priors carefully (actually not too much of an issue)
  - Computation (e.g. Stan)

Experiments Setup

- 29 LLMs
    - actually have ~60
- 24 Tasks (BigBench Hard) (about 200-250 questions each)
    - actually have at least ~40 tasks
- Raw eval binary data (success/fail) from Huggingface leaderboard $\mathcal{D}$

Open LLM Leaderboard

Comparing Large Language Models in an **open** and **reproducible** way

```
meta-llama/Meta-Llama-3.1-70B
meta-llama/Meta-Llama-3.1-8B
meta-llama/Meta-Llama-3-8B
meta-llama/Meta-Llama-3-70B
mistralai/Mixtral-8x7B-v0.1
mistralai/Mixtral-8x22B-v0.1
google/gemma-2-27b
google/gemma-2-2b
Qwen/Qwen1.5-7B
Qwen/Qwen1.5-110B
Qwen/Qwen2.5-0.5B
Qwen/Qwen2.5-3B
Qwen/Qwen2.5-14B
Qwen/Qwen2.5-72B
meta-llama/Meta-Llama-3.1-70B-Instruct
meta-llama/Meta-Llama-3.1-8B-Instruct
meta-llama/Meta-Llama-3-8B-Instruct
meta-llama/Meta-Llama-3-70B-Instruct
microsoft/Phi-3.5-mini-instruct
microsoft/Phi-3.5-MoE-instruct
mistralai/Mixtral-8x7B-Instruct-v0.1
mistralai/Mistral-7B-Instruct-v0.3
google/gemma-2-27b-it
Qwen/Qwen1.5-7B-Chat
Qwen/Qwen1.5-110B-Chat
Qwen/Qwen2.5-0.5B-Instruct
Qwen/Qwen2.5-3B-Instruct
Qwen/Qwen2.5-14B-Instruct
Qwen/Qwen2.5-72B-Instruct
```

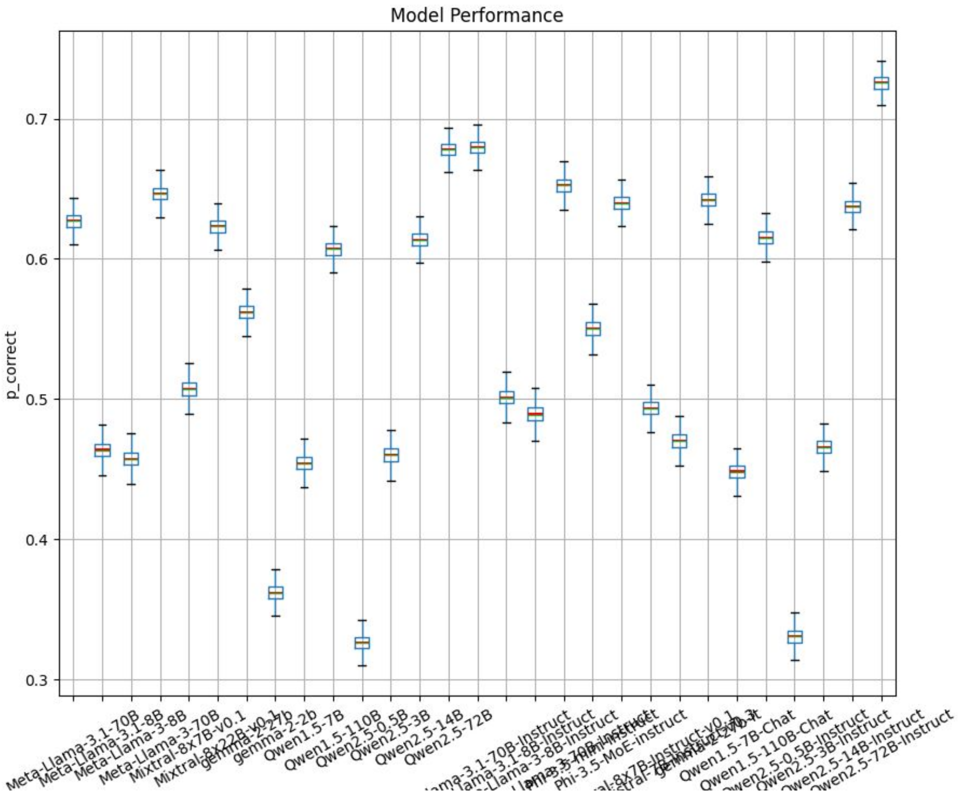
```
"bbh_boolean_expressions",
"bbh_causal_judgement",
"bbh_date_understanding",
"bbh_disambiguation_qa",
"bbh_formal_fallacies",
"bbh_geometric_shapes",
"bbh_hyperbaton",
"bbh_logical_deduction_five_objects",
"bbh_logical_deduction_seven_objects",
"bbh_logical_deduction_three_objects",
"bbh_movie_recommendation",
"bbh_navigate",
"bbh_object_counting",
"bbh_penguins_in_a_table",
"bbh_reasoning_about_colored_objects",
"bbh_ruin_names",
"bbh_salient_translation_error_detection",
"bbh_snarks",
"bbh_sports_understanding",
"bbh_temporal_sequences",
"bbh_tracking_shuffled_objects_five_objects",
"bbh_tracking_shuffled_objects_seven_objects",
"bbh_tracking_shuffled_objects_three_objects",
"bbh_web_of_lies",
```

Model 1: Beta Binomial

Prior $p_m \sim \mathrm{Beta}(1, 1)$

Model $\bar{s}_{m,t} \sim \mathrm{Bin}(N_t, p_m)$

Infer $\{p_m\}_{m=1}^{M} | \mathcal{D}$

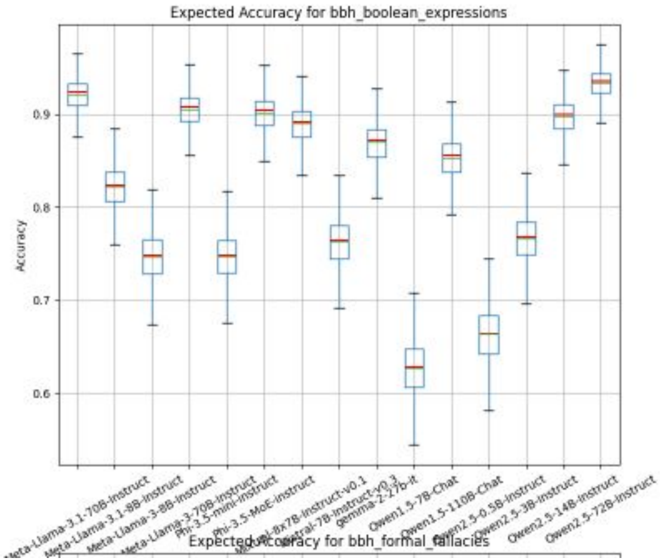


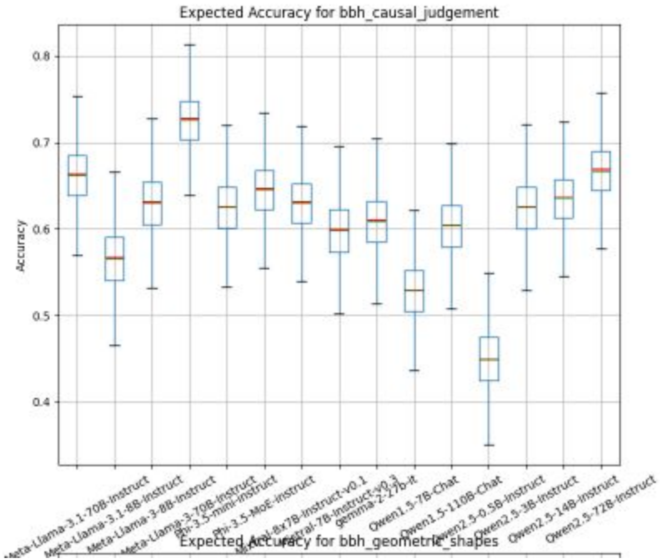
Model 2: Beta Binomial with per-task model performance

Prior $p_{m,t} \sim \mathrm{Beta}(1,1)$

Model $\bar{s}_{m,t} \sim \mathrm{Bin}(N_t, p_{m,t})$

Infer $\{p_{m,t}\}_{m=1}^{M} | \mathcal{D}$



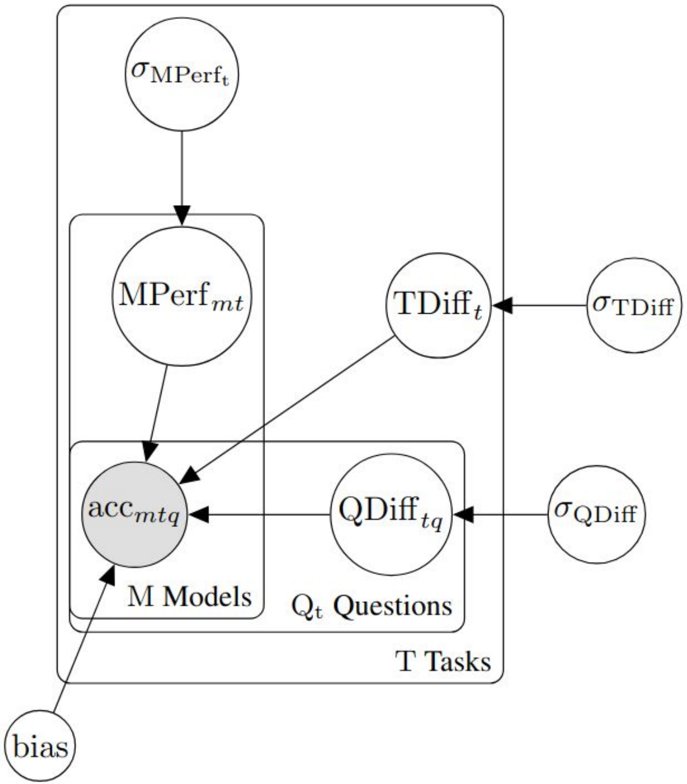


Model 3: Question-based model

$$\text{bias} \sim \mathcal{N}(0,1)$$
$$\text{taskDifficultyStd} \sim \text{Gamma}(2,2)$$
$$\text{questionDifficultyStd} \sim \text{Gamma}(2,2)$$
$$\text{modelPerformanceStd} \sim \text{Gamma}(2,2)$$

$$\text{taskDifficulty}_t \sim \mathcal{N}(0, \text{taskDifficultyStd})$$
$$\text{questionDifficulty}_q \sim \mathcal{N}(0, \text{questionDifficultyStd})$$
$$\text{modelPerformance}_t \sim \mathcal{N}(0, \text{modelPerformanceStd})$$

$$\begin{aligned}\bar{s}_{m,q} \sim \text{Ber}(\text{sigmoid}(\text{bias} \\ &+ \text{modelPerformance}_m \\ &- \text{taskDifficulty}_{\text{task}(q)} \\ &- \text{questionDifficulty}_q))\end{aligned}$$

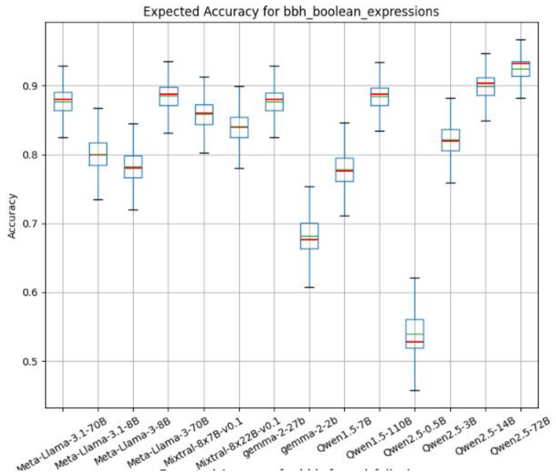


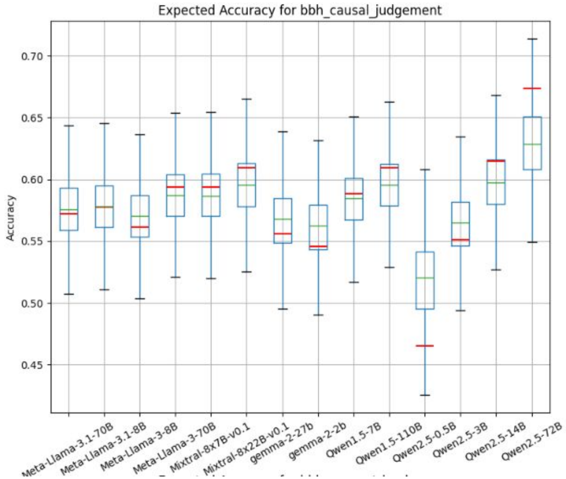
Model 3: Question-based model

$$\text{bias} \sim \mathcal{N}(0, 1)$$
$$\text{taskDifficultyStd} \sim \text{Gamma}(2, 2)$$
$$\text{questionDifficultyStd} \sim \text{Gamma}(2, 2)$$
$$\text{modelPerformanceStd} \sim \text{Gamma}(2, 2)$$

$$\text{taskDifficulty}_t \sim \mathcal{N}(0, \text{taskDifficultyStd})$$
$$\text{questionDifficulty}_q \sim \mathcal{N}(0, \text{questionDifficultyStd})$$
$$\text{modelPerformance}_t \sim \mathcal{N}(0, \text{modelPerformanceStd})$$

$$\begin{aligned}\bar{s}_{m,q} \sim \text{Ber}(\text{sigmoid}(\text{bias} & \\ & + \text{modelPerformance}_m \\ & - \text{taskDifficulty}_{\text{task}(q)} \\ & - \text{questionDifficulty}_q))\end{aligned}$$





Model 4: Question-based model w/ across-task performance

$$\text{bias} \sim \mathcal{N}(0, 1)$$
$$\text{taskDifficultyStd} \sim \text{Gamma}(2, 2)$$
$$\text{questionDifficultyStd} \sim \text{Gamma}(2, 2)$$
$$\text{modelPerformanceStd} \sim \text{Gamma}(2, 2)$$
$$\text{acrossTaskPerformanceStd} \sim \text{Gamma}(2, 2)$$
$$\text{taskDifficulty}_t \sim \mathcal{N}(0, \text{taskDifficultyStd})$$
$$\text{questionDifficulty}_q \sim \mathcal{N}(0, \text{questionDifficultyStd})$$
$$\text{modelPerformance}_{m,t} \sim \mathcal{N}(0, \text{modelPerformanceStd})$$
$$\text{acrossTaskPerformance}_m \sim \mathcal{N}(0, \text{acrossTaskPerformanceStd})$$

$$\bar{s}_{m,q} \sim \text{Ber}(\text{sigmoid}(\text{bias} + \text{modelPerformance}_{m,t} + \text{acrossTaskPerformance}_m - \text{taskDifficulty}_{\text{task}(q)} - \text{questionDifficulty}_q))$$

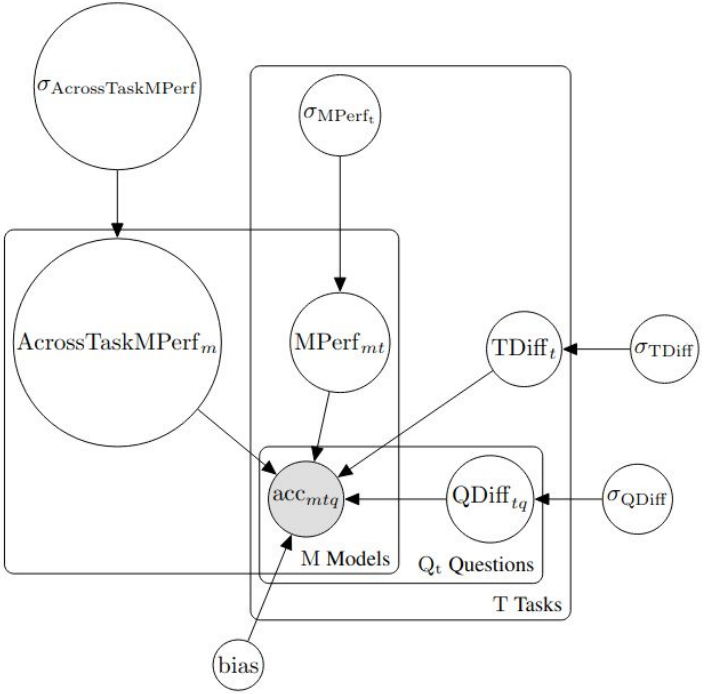
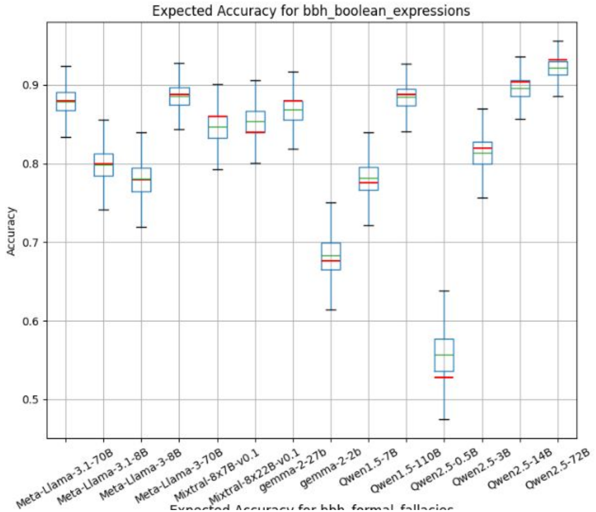


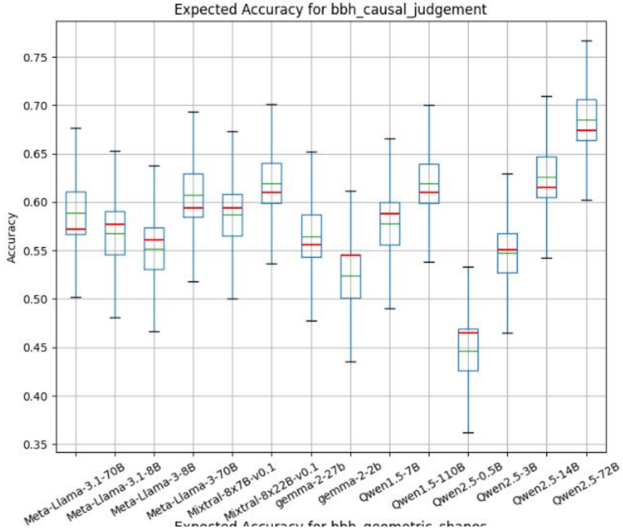
Figure 4. Graphical model for the hierarchical model with per-model overall performance variables.

Model 3: w/ across-task performance

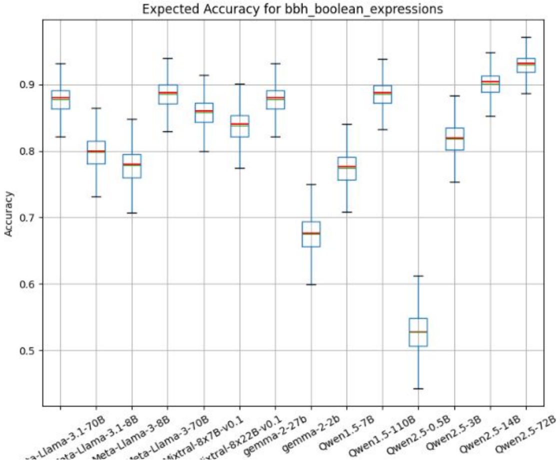
$$\text{bias} \sim \mathcal{N}(0, 1)$$
$$\text{taskDifficultyStd} \sim \text{Gamma}(2, 2)$$
$$\text{questionDifficultyStd} \sim \text{Gamma}(2, 2)$$
$$\text{modelPerformanceStd} \sim \text{Gamma}(2, 2)$$
$$\text{acrossTaskPerformanceStd} \sim \text{Gamma}(2, 2)$$
$$\text{taskDifficulty}_t \sim \mathcal{N}(0, \text{taskDifficultyStd})$$
$$\text{questionDifficulty}_q \sim \mathcal{N}(0, \text{questionDifficultyStd})$$
$$\text{modelPerformance}_{m,t} \sim \mathcal{N}(0, \text{modelPerformanceStd})$$
$$\text{acrossTaskPerformance}_m \sim \mathcal{N}(0, \text{acrossTaskPerformanceStd})$$

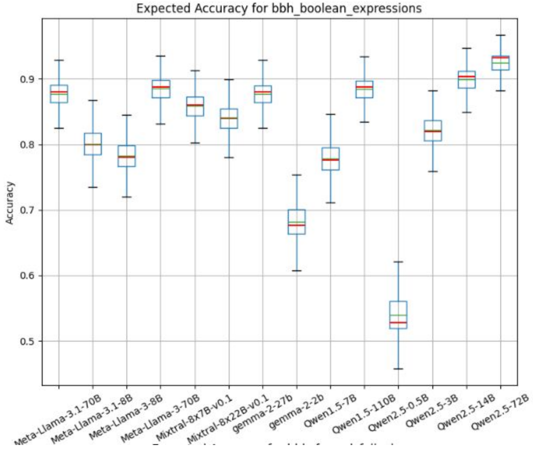
$$\begin{aligned}\bar{s}_{m,q} \sim \text{Ber}(\text{sigmoid}(\text{bias} & \\ & + \text{modelPerformance}_{m,t} \\ & + \text{acrossTaskPerformance}_m \\ & - \text{taskDifficulty}_{\text{task}(q)} \\ & - \text{questionDifficulty}_q))\end{aligned}$$

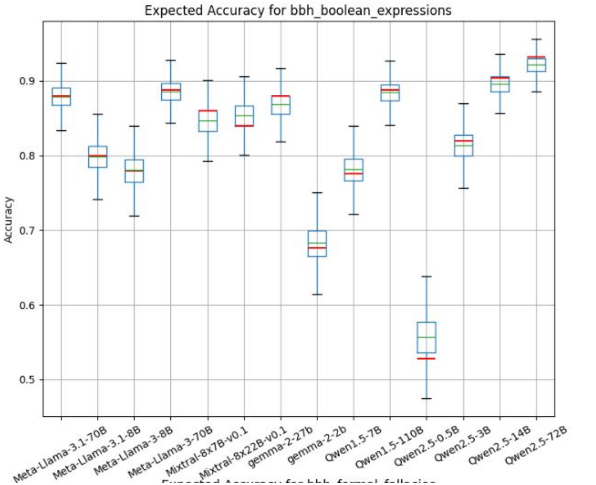


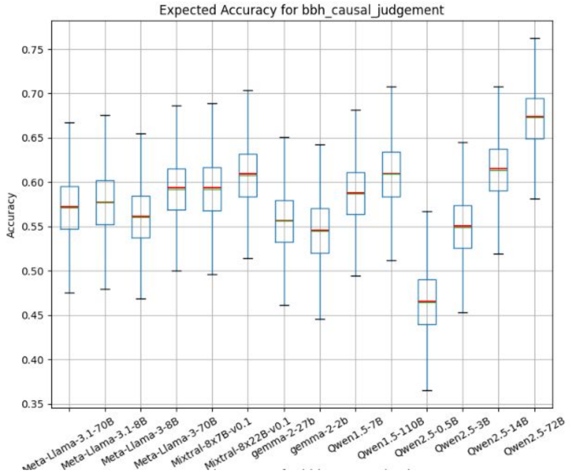


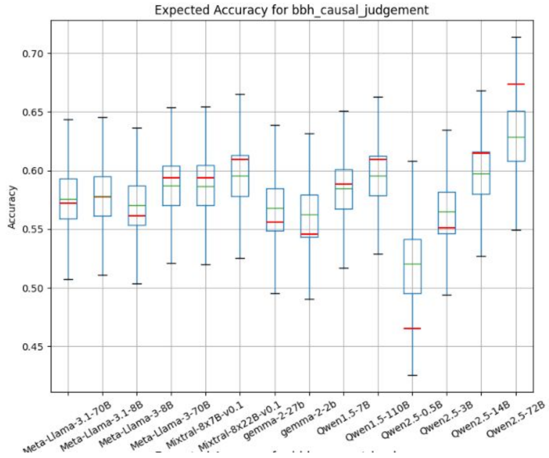
Model 2 vs 3 vs 4 - some (slight) reduction in size of error bars – needs frequentist comparison

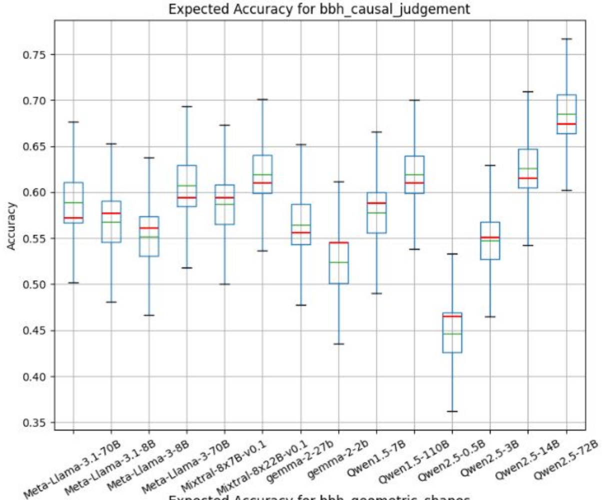












Bayesian hierarchical modelling for evals makes it easy to…

1. Easily find a Prob(Model A > Model B) – HMC gives us samples of $p_{m,t}|\mathcal{D}$, just look at

$$\frac{1}{\text{num samples}} \sum_{\text{samples}} \frac{1}{T} \sum_{t} \mathbb{I}[p_{A,t} > p_{B,t}]$$

Bayesian hierarchical modelling for evals makes it easy to…

1. Easily find a Prob(Model A > Model B) – HMC gives us samples of $p_{m,t}|\mathcal{D}$, just look at

$$\frac{1}{\text{num samples}} \sum_{\text{samples}} \frac{1}{T} \sum_{t} \mathbb{I}[p_{A,t} > p_{B,t}]$$

2. Elicit information about latent variables of interest:
    a. Task difficulty
    b. Per-task model performance
    c. Capabilities*

Bayesian hierarchical modelling for evals makes it easy to…

1. Easily find a Prob(Model A > Model B) – HMC gives us samples of $p_{m,t}|\mathcal{D}$, just look at

$$\frac{1}{\text{num samples}} \sum_{\text{samples}} \frac{1}{T} \sum_t \mathbb{I}[p_{A,t} > p_{B,t}]$$

2. Elicit information about latent variables of interest:
    a. Task difficulty
    b. Per-task model performance
    c. Capabilities*

* If we make modelPerformance and taskDifficulty vectors of length C (say 5 or 7) with likelihood s.t. $\bar{s}_{m,t}$ depends on their dot product, each dimension of those vectors might be interpretable as model/task capabilities e.g. arithmetic, grammar, logic etc.

References

- Anthropic. 2023. ‘Challenges in Evaluating AI Systems’. 4 October 2023. https://www.anthropic.com/news/evaluating-ai-systems.
- Biderman, Stella, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, et al. 2024. ‘Lessons from the Trenches on Reproducible Evaluation of Language Models’. arXiv. https://doi.org/10.48550/arXiv.2405.14782.
- Grattafiori, Aaron, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, et al. 2024. ‘The Llama 3 Herd of Models’. arXiv. https://doi.org/10.48550/arXiv.2407.21783.
- Gu, Yuling, Oyvind Tafjord, Bailey Kuehl, Dany Haddad, Jesse Dodge, and Hannaneh Hajishirzi. 2024. ‘OLMES: A Standard for Language Model Evaluations’. arXiv. https://doi.org/10.48550/arXiv.2406.08446.
- Hermann, Katherine, Jennifer Hu, and Michael Mozer. n.d. ‘Experimental Design and Analysis for AI Researchers’.
- Ivanova, Desi. 2024a. ‘On Some (Fixable) Limitations of “Understanding the Limitations of Mathematical Reasoning in LLMs”’. 22 October 2024. https://substack.com/inbox/post/150508215?utm_campaign=post&triedRedirect=true.
- ———. 2024b. ‘Towards More Rigorous Evaluations of Language Models’. 28 November 2024. https://substack.com/home/post/p-152149873.
- Liang, Percy, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, et al. 2023. ‘Holistic Evaluation of Language Models’. arXiv. https://doi.org/10.48550/arXiv.2211.09110.
- Miller, Evan. 2024. ‘Adding Error Bars to Evals: A Statistical Approach to Language Model Evaluations’. arXiv. https://doi.org/10.48550/arXiv.2411.00640.
- Mirzadeh, Iman, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2024. ‘GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models’. arXiv. https://doi.org/10.48550/arXiv.2410.05229.
- Perez, Ethan, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, et al. 2022. ‘Discovering Language Model Behaviors with Model-Written Evaluations’. arXiv. https://doi.org/10.48550/arXiv.2212.09251.
- Polo, Felipe Maia, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. 2024. ‘tinyBenchmarks: Evaluating LLMs with Fewer Examples’. arXiv. https://doi.org/10.48550/arXiv.2402.14992.
- Reuel, Anka, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J. Kochenderfer. 2024. ‘BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices’. arXiv. https://doi.org/10.48550/arXiv.2411.12990.
- Song, Yifan, Guoyin Wang, Sujian Li, and Bill Yuchen Lin. 2024. ‘The Good, The Bad, and The Greedy: Evaluation of LLMs Should Not Ignore Non-Determinism’. arXiv. https://doi.org/10.48550/arXiv.2407.10457.