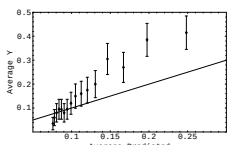
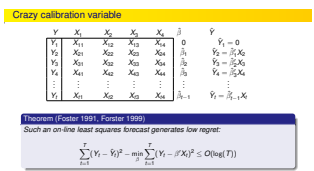
Even LLMs can be calibrated

Dean Foster

Amazon & U Penn

November 2, 2026





What is the chance of rain tomorrow?

What is the chance of rain tomorrow?

. . . the chance the LLM answered correctly?

pred	out	pred	out	pred	out	pred	out	pred	out
.3	✔	.9	✔	.6	✘	.1	✘	.8	✔
.8	✘	.3	✘	.2	✔	.3	✘	.8	✘
.4	✔	.5	✔	.8	✘	.3	✘	.3	✔
.0	✘	.3	✘	.4	✔	.6	✘	.8	✘
.3	✘	.9	✔	.9	✔	.2	✔	.4	✔
.5	✔	.0	✘	.0	✘	.1	✘	.3	✔
.6	✘	.3	✘	.3	✔	.3	✘	.9	✔
.2	✔	.5	✔	.5	✔	.1	✘	.0	✘
.3	✘	.0	✘	.8	✔	.3	✘	.4	✘
.9	✔	.3	✘	.3	✘	.3	✔	.6	✘
.3	✘	.5	✔	.4	✔	.8	✘	.2	✔
.4	✔	.3	✘	.3	✔	.3	✘	.8	✔

pred	out	pred	out	pred	out	pred	out	pred	out
.3	✔	.9	✔	.6	✘	.1	✘	.8	✔
.8	✘	.3	✘	.2	✔	.3	✘	.8	✘
.4	✔	.5	✔	.8	✘	.3	✘	.3	✔
.0	✘	.3	✘	.4	✔	.6	✘	.8	✘
.3	✘	.9	✔	.9	✔	.2	✔	.4	✔
.5	✔	.0	✘	.0	✘	.1	✘	.3	✔
.6	✘	.3	✘	.3	✔	.3	✘	.9	✔
.2	✔	.5	✔	.5	✔	.1	✘	.0	✘
.3	✘	.0	✘	.8	✔	.3	✘	.4	✘
.9	✔	.3	✘	.3	✘	.3	✔	.6	✘
.3	✘	.5	✔	.4	✔	.8	✘	.2	✔
.4	✔	.3	✘	.3	✔	.3	✘	.8	✔

pred	out	pred	out
.3	✔	.3	✘
.3	✘	.3	✔
.3	✘	.3	✘
.3	✘	.3	✘
.3	✘	.3	✘
.3	✘	.3	✘
.3	✘	.3	✔
.3	✘	.3	✘
.3	✘	.3	✔
.3	✔	.3	✔

$$.3 = \frac{6✔}{6✔ + 14✘}$$

pred	out	pred	out
.3	✔	.3	✘
.3	✘	.3	✔
.3	✘	.3	✘
.3	✘	.3	✘
.3	✘	.3	✘
.3	✘	.3	✘
.3	✘	.3	✔
.3	✘	.3	✘
.3	✘	.3	✔
.3	✔	.3	✔

$$.3 = \frac{6✔}{6✔ + 14✘}$$

perfect calibration!

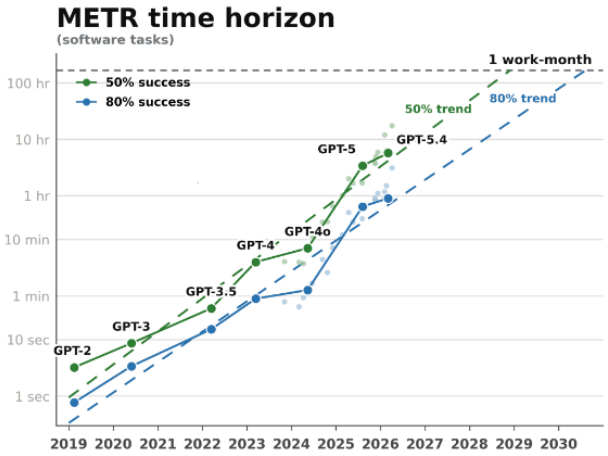
I've been working on calibration a long time

- ~~1992~~ 1998: Asymptotic Calibration.
- 1997: Calibrated Learning and Correlated Equilibrium.
- 1999: A proof of Calibration via Blackwell's Approachability Theorem.
- 2006: Calibration via Regression.
- 2008: Deterministic Calibration and Nash Equilibrium.
- 2011: Complexity-Based Approach to Calibration with Checking Rules.
- 2012: Calibration: Respice, Adspice, Prospice.
- 2018: Smooth calibration, leaky forecasts, finite recall, and Nash dynamics.
- 2021: Forecast Hedging and Calibration.
- 2026: Calibeating: Beating forecasters at their own game

How often is an LLM correct?

- Modelling the accuracy of an LLM answering a questions isn't easy
- Typically the questions are unrelated questions
- We can't asume IID, stationary, ergodic, etc
- So we need to work with individual sequences

LLMs are not stationary (source: Halperin)



Crazy calibration variable

Y	X_1	X_2	X_3	X_4
Y_1	X_{11}	X_{12}	X_{13}	X_{14}
Y_2	X_{21}	X_{22}	X_{23}	X_{24}
Y_3	X_{31}	X_{32}	X_{33}	X_{34}
Y_4	X_{41}	X_{42}	X_{43}	X_{44}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y_t	X_{t1}	X_{t2}	X_{t3}	X_{t4}

Starting with our data that we observed up to time t

Crazy calibration variable

Y	X_1	X_2	X_3	X_4
Y_1	X_{11}	X_{12}	X_{13}	X_{14}
Y_2	X_{21}	X_{22}	X_{23}	X_{24}
Y_3	X_{31}	X_{32}	X_{33}	X_{34}
Y_4	X_{41}	X_{42}	X_{43}	X_{44}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y_t	X_{t1}	X_{t2}	X_{t3}	X_{t4}

$$\hat{\beta}_t = \arg\min_\beta \sum_{i=1}^{t} (Y_i - \beta' X_i)^2$$

We can fit $\hat{\beta}_t$ on everything up to time t

Crazy calibration variable

Y	X_1	X_2	X_3	X_4
Y_1	X_{11}	X_{12}	X_{13}	X_{14}
Y_2	X_{21}	X_{22}	X_{23}	X_{24}
Y_3	X_{31}	X_{32}	X_{33}	X_{34}
Y_4	X_{41}	X_{42}	X_{43}	X_{44}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y_t	X_{t1}	X_{t2}	X_{t3}	X_{t4}
	$X_{t+1,1}$	$X_{t+1,2}$	$X_{t+1,3}$	$X_{t+1,4}$

$\hat{\beta}_t$

$$\hat{Y}_{t+1} = \hat{\beta}_t' X_{t+1}$$

From a new X_{t+1} we can compute $\hat{Y}_{t+1}$

Crazy calibration variable

Y	X_1	X_2	X_3	X_4	$\hat{\beta}$
Y_1	X_{11}	X_{12}	X_{13}	X_{14}	0
Y_2	X_{21}	X_{22}	X_{23}	X_{24}	$\hat{\beta}_1$
Y_3	X_{31}	X_{32}	X_{33}	X_{34}	$\hat{\beta}_2$
Y_4	X_{41}	X_{42}	X_{43}	X_{44}	$\hat{\beta}_3$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y_t	X_{t1}	X_{t2}	X_{t3}	X_{t4}	$\hat{\beta}_{t-1}$

Looking at only the first part of the data, we can generate:

$$\hat{\beta}_0, \quad \hat{\beta}_1, \quad \hat{\beta}_2, \quad \hat{\beta}_3, \quad \hat{\beta}_4, \quad \ldots, \quad \hat{\beta}_{t-1}$$

Crazy calibration variable

Y	X_1	X_2	X_3	X_4	$\hat{\beta}$	$\hat{Y}$
Y_1	X_{11}	X_{12}	X_{13}	X_{14}	0	$\hat{Y}_1 = 0$
Y_2	X_{21}	X_{22}	X_{23}	X_{24}	$\hat{\beta}_1$	$\hat{Y}_2 = \hat{\beta}_1' X_2$
Y_3	X_{31}	X_{32}	X_{33}	X_{34}	$\hat{\beta}_2$	$\hat{Y}_3 = \hat{\beta}_2' X_3$
Y_4	X_{41}	X_{42}	X_{43}	X_{44}	$\hat{\beta}_3$	$\hat{Y}_4 = \hat{\beta}_3' X_4$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y_t	X_{t1}	X_{t2}	X_{t3}	X_{t4}	$\hat{\beta}_{t-1}$	$\hat{Y}_t = \hat{\beta}_{t-1}' X_t$

Each of these leads to a next round

$$\hat{Y}_1, \quad \hat{Y}_2, \quad \hat{Y}_3, \quad \hat{Y}_4, \quad \ldots, \quad \hat{Y}_t$$

Crazy calibration variable

Y	X_1	X_2	X_3	X_4	$\hat{\beta}$	$\hat{Y}$
Y_1	X_{11}	X_{12}	X_{13}	X_{14}	0	$\hat{Y}_1 = 0$
Y_2	X_{21}	X_{22}	X_{23}	X_{24}	$\hat{\beta}_1$	$\hat{Y}_2 = \hat{\beta}_1' X_2$
Y_3	X_{31}	X_{32}	X_{33}	X_{34}	$\hat{\beta}_2$	$\hat{Y}_3 = \hat{\beta}_2' X_3$
Y_4	X_{41}	X_{42}	X_{43}	X_{44}	$\hat{\beta}_3$	$\hat{Y}_4 = \hat{\beta}_3' X_4$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y_t	X_{t1}	X_{t2}	X_{t3}	X_{t4}	$\hat{\beta}_{t-1}$	$\hat{Y}_t = \hat{\beta}_{t-1}' X_t$

Theorem (Foster 1991, Forster 1999)

Such an on-line least squares forecast generates low regret:

$$\sum_{t=1}^{T} (Y_t - \hat{Y}_t)^2 - \min_{\beta} \sum_{t=1}^{T} (Y_t - \beta' X_t)^2 \leq O(\log(T))$$

Crazy calibration variable

Y	X_1	X_2	X_3	X_4	$\hat{\beta}$	$\hat{Y}$
Y_1	X_{11}	X_{12}	X_{13}	X_{14}	0	$\hat{Y}_1 = 0$
Y_2	X_{21}	X_{22}	X_{23}	X_{24}	$\hat{\beta}_1$	$\hat{Y}_2 = \hat{\beta}_1' X_2$
Y_3	X_{31}	X_{32}	X_{33}	X_{34}	$\hat{\beta}_2$	$\hat{Y}_3 = \hat{\beta}_2' X_3$
Y_4	X_{41}	X_{42}	X_{43}	X_{44}	$\hat{\beta}_3$	$\hat{Y}_4 = \hat{\beta}_3' X_4$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y_t	X_{t1}	X_{t2}	X_{t3}	X_{t4}	$\hat{\beta}_{t-1}$	$\hat{Y}_t = \hat{\beta}_{t-1}' X_t$

Works no matter what the X's are.

Crazy calibration variable

Y	X_1	X_2	X_3	X_4	$\hat{\beta}$	$\hat{Y}$
Y_1	X_{11}	X_{12}	$\hat{Y}_1$	X_{14}	0	$\hat{Y}_1 = 0$
Y_2	X_{21}	X_{22}	$\hat{Y}_2$	X_{24}	$\hat{\beta}_1$	$\hat{Y}_2 = \hat{\beta}_1' X_2$
Y_3	X_{31}	X_{32}	$\hat{Y}_3$	X_{34}	$\hat{\beta}_2$	$\hat{Y}_3 = \hat{\beta}_2' X_3$
Y_4	X_{41}	X_{42}	$\hat{Y}_4$	X_{44}	$\hat{\beta}_3$	$\hat{Y}_4 = \hat{\beta}_3' X_4$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y_t	X_{t1}	X_{t2}	$\hat{Y}_t$	X_{t4}	$\hat{\beta}_{t-1}$	$\hat{Y}_t = \hat{\beta}_{t-1}' X_t$

Even if one of the X's were $\hat{Y}$!

Crazy calibration variable

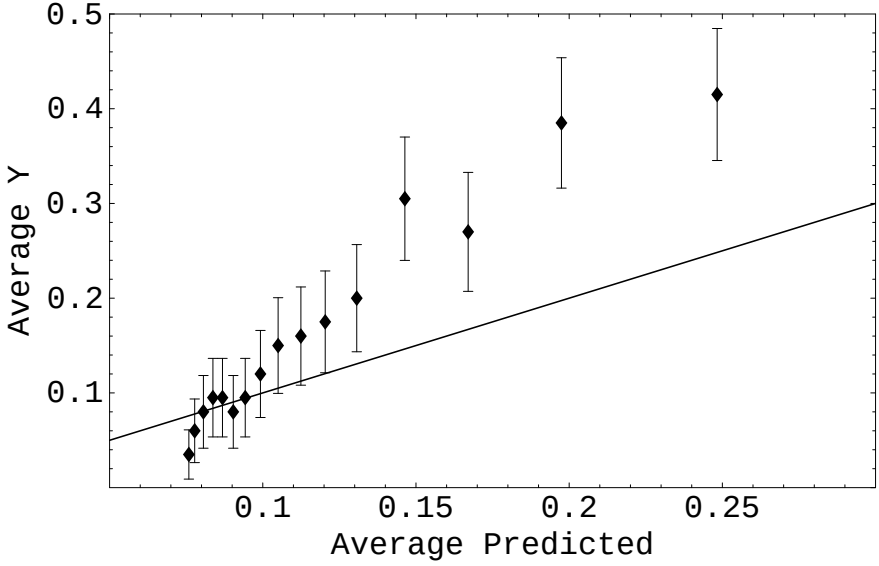
Y	X_1	X_2	X_3	X_4	$\hat{\beta}$	$\hat{Y}$
Y_1	X_{11}	X_{12}	$\hat{Y}_1$	X_{14}	0	$\hat{Y}_1 = 0$
Y_2	X_{21}	X_{22}	$\hat{Y}_2$	X_{24}	$\hat{\beta}_1$	$\hat{Y}_2 = \hat{\beta}_1' X_2$
Y_3	X_{31}	X_{32}	$\hat{Y}_3$	X_{34}	$\hat{\beta}_2$	$\hat{Y}_3 = \hat{\beta}_2' X_3$
Y_4	X_{41}	X_{42}	$\hat{Y}_4$	X_{44}	$\hat{\beta}_3$	$\hat{Y}_4 = \hat{\beta}_3' X_4$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y_t	X_{t1}	X_{t2}	$\hat{Y}_t$	X_{t4}	$\hat{\beta}_{t-1}$	$\hat{Y}_t = \hat{\beta}_{t-1}' X_t$

Theorem ($\Longrightarrow$ Foster and Kakade 2008, Foster and Hart 2018)

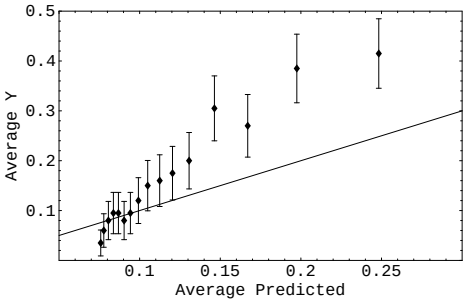
Adding the crazy calibration variable gives multi-calibration:

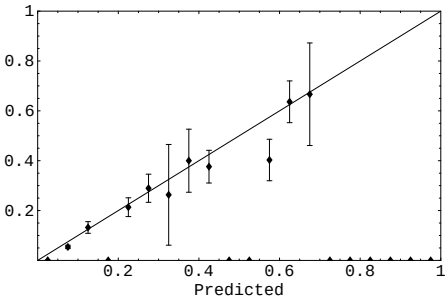
$$(\forall i) \quad \sum_{t=1}^{T} X_{t,i}(Y_t - \hat{Y}_t) = O(\sqrt{T \log(T)})$$

This is not calibrated



Anything easily fixed isn't calibrated





Fix the obvious problems!

Calibration limits LLM misalignment

- Calibration is a low bar

Calibration limits LLM misalignment

- Calibration is a low bar
- We want more: Multi-calibration
  - Suppose we want to be calibrated on men and women
  - We add a indicator function
  - Now we are unbiased on both subpopulation
  - Just like regular regression!
  - All due to the normal equation holding

Multi-calibration is the protection we need

If we use multi-calibration, we are calibrated on:

- mathematics
- history
- factual questions
- theorems
- trivial questions
- meaningful questions

Multi-calibration is the protection we need

If we use multi-calibration, we are calibrated on:

- mathematics
- history
- factual questions
- theorems
- trivial questions
- meaningful questions

(But not necessarily on historical$\times$factual$\times$trivial questions)

Same model
same question
using chain-of-thought changes accuracy

Contrasting native LLMs vs Chain of Thought

Theorem (Merrill and Sabharwal 2023)

(rephrased) An LLM cannot answer questions in PSPACE. [arXiv]

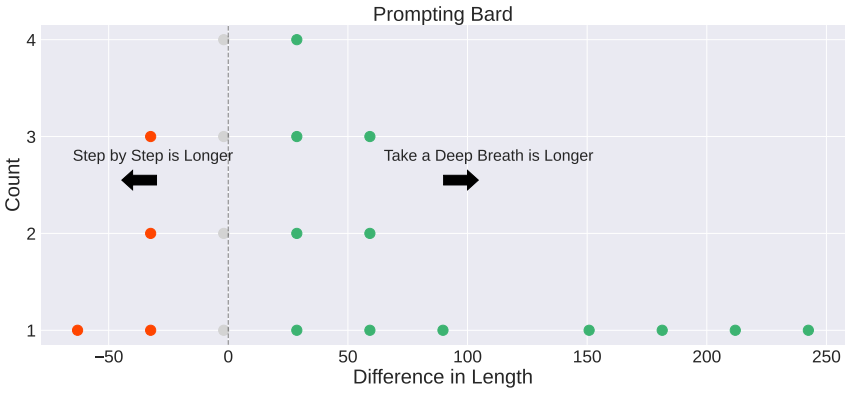
Theorem (F. and Madeka 2023)

Using chain of thought reasoning, an LLM can solve any problem in PSPACE.

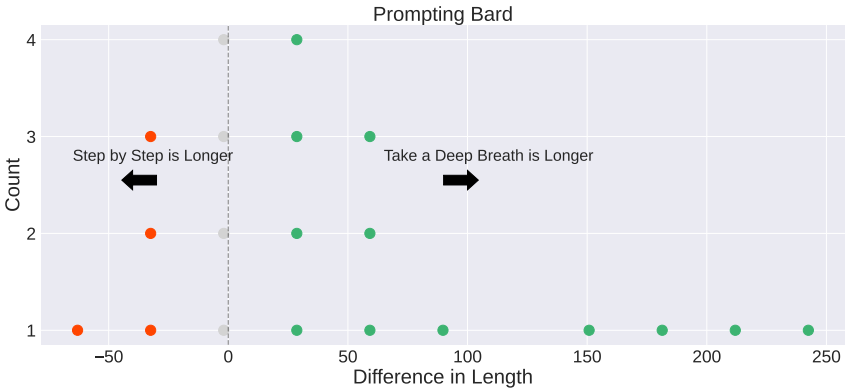
[slides]

Simply changing the prompt matters!

LLMs do better with more thinking



LLMs do better with more thinking

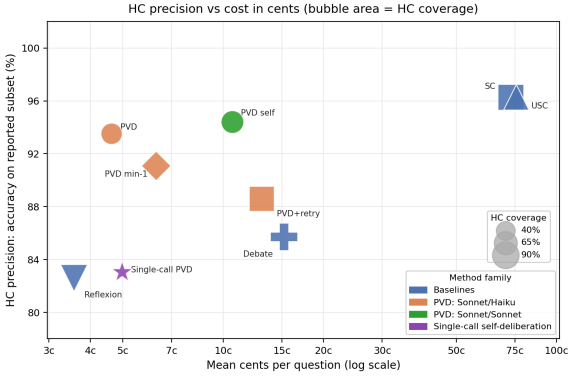


(This year: “I believe in you”)

More than one LLM

Prover–Verifier Deliberation

- Prover defends an answer through checkable sub-claims; verifier returns ACCEPT, CHALLENGE, or REJECT.
- 30% difference between the Accept and Reject endpoints



More than one forecaster

Calibeating

Definition

c calibeats *b* if it is more accurate than a calibrated version of *b*.

Calibeating

Definition

c calibeats *b* if it is more accurate than a calibrated version of *b*.

- x_t is our "rain" / "no rain"
- b_t is our *bad* forecast
- c_t is a *calibeating* forecast
- $a_t = A(b_t, c_t)$ is an action
- $\mathcal{L}(a, x)$ is a loss function

$$\sum_{t=1}^{T} \mathcal{L}(a_t, x_t) = \sum_{t=1}^{T} E_{X \sim c_t}(\mathcal{L}(a_t, X)) + o(T)$$

- LHS talks about arbitrary sequences
- RHS talks about easy optimizations

Standard errors

- What does a SE(b_t) mean?
  - frequents need a model
  - Bayesians consider it non-cognitivism
  - What does a SE look like for calibration?

Standard errors

- What does a SE(b_t) mean?
  - frequents need a model
  - Bayesians consider it non-cognitivism
  - What does a SE look like for calibration?
- SE(b_t) is a calibrated forecast of:

$$|b_t - c_t|$$

Swarms

- Ask agents for calibrated estimates and calibrated SE.
  - If the SE is large, keep pestering them for more deliberation and chain of thought.
  - If the SE is zero, stop and save money
- Show them each other's work
  - use deliberation and iterated calibeating
- Ask cheap agents first
  - maybe the problem is easy and you won't have to burn tokens on expensive agents.

Summary

LLMs change each day. IID is not a good model for them.

Summary

LLMs change each day. IID is not a good model for them.

Four concepts which are a good model for LLMs:

- Calibration gets the average right

Summary

LLMs change each day. IID is not a good model for them.

Four concepts which are a good model for LLMs:

- Calibration gets the average right
- Multicalibration: gets sub-populations right

Summary

LLMs change each day. IID is not a good model for them.

Four concepts which are a good model for LLMs:

- Calibration gets the average right
- Multicalibration: gets sub-populations right
- Calibeating supposts actions and agent interactions

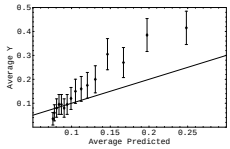
Summary

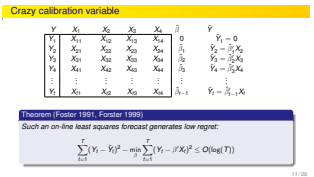
LLMs change each day. IID is not a good model for them.

Four concepts which are a good model for LLMs:

- Calibration gets the average right
- Multicalibration: gets sub-populations right
- Calibeating supposts actions and agent interactions
- Calibrated SEs are meaningful and useful

THANKS!





(slides are at deanfoster.net)

Citations (slides at deanfoster.net)

"Running Out Of GDP" Basil Halperin (2026).

"A TCS look at LLMs." [5]

"Trust but Verify: Prover–Verifier Deliberation for Selective LLM Prediction," Sedoc, Zhang, F. [7]

"Human–LLM Deliberation as an Interactive Zero-Knowledge Proof Protocol," Sedoc et al.

"Calibeating: Beating Forecasters at Their Own Game," F. and Hart. [8]

"Tractable Agreement Protocols," Collina, Goel, Gupta, Roth. [arXiv]

"L3M coding harness" an open source harness which uses deliberation and calibration.