

Closed-book overfitting (multiple config and metric testing)

Part of *Understanding and Mastering Uncertainty*

Many ways to Impress Your Boss

PICK THE METRIC

Dataset and configuration fixed

Metrics

PICK THE CONFIGURATION

Dataset and metric fixed

Configurations

PICK THE DATASET

Configuration and metric fixed

Test dataset



Every match sets a world record!

Southampton 1 – 1 Sunderland.

A forgettable Tuesday night.

The commentator still has this ready for the 88th minute:

RECORD

First left-footed header scored on a Tuesday in March since records began in 2003

RECORD

Longest spell without a corner in a match with exactly two yellow cards

None of these was a question **before** kick-off.

The record was written **after** the match, by whoever looked hardest.

There is always a record

90

minutes to pick
from

×

40

things to count:
passes, shots, corners,
saves, fouls...

×

200

ways to slice:
name starts with S,
red kit, on a
Tuesday...

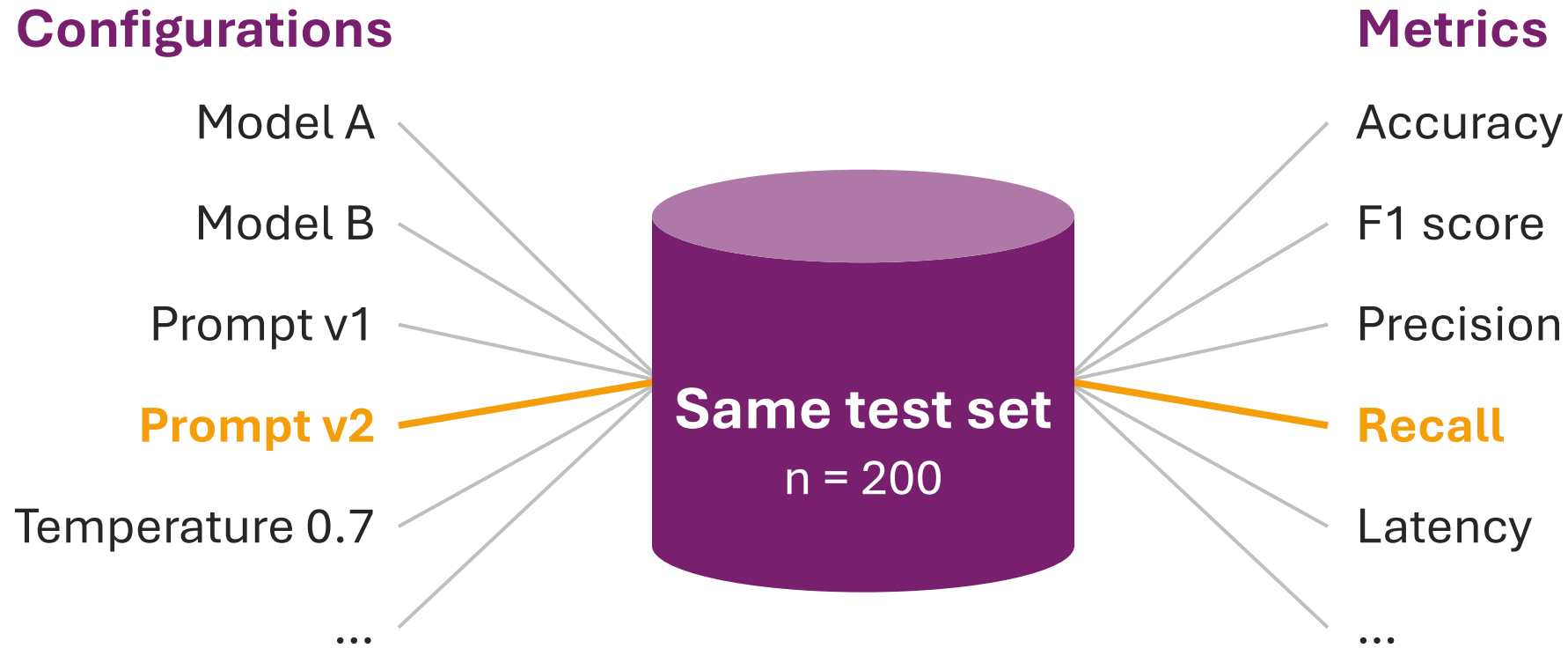
×

4

ways to be a
record: most,
fewest, first, longest

= 2,880,000 “candidate records” in one ordinary match.

One test set, many questions



Every line is a “question” asked of the same data.

Ask enough, and one looks good by chance.

Two types of closed-book overfitting

1. We can “find” stories that do not exist ([story playground](#))
2. We inflate estimates ([inflated estimates playground](#))

And: [Playground](#): upload your metrics, see what you can conclude

Story Finder

Look at enough numbers and a story will find you. Explore → Claim → Challenge → Compare.

►What is this playground about? (click to expand)

Explore: find stories, then challenge them

Pre-register a comparison instead

Reality: there is no story.

Every model and every condition has **exactly the same true mean and spread, on every slice**; cells differ only in which 200 tickets they happened to draw. "Significant" means $p < 0.05$: a difference this big shows up in **1 of 20 samples even when the two cells are truly identical**, so 660 comparisons produce about 33 "significant" differences on average even if nothing is real. A single run can show far fewer or far more, because every comparison shares the same baseline sample.

Start here: choose which conditions to vary and watch the number of cells, tokens and *looks* multiply. Then click "Run experiment" (and again, and again).

EXPERIMENTAL CONDITIONS (what you vary, and on how many tickets)

- ☒ **Model**
Model A, Model B, Model C
- ☒ **Prompt**
Prompt variant 1, Prompt variant 2, Prompt variant 3, Prompt variant 4
- ☐ **Temperature**
T=0.0, T=0.7
- ☐ **Dataset**
Retail, Telco, Banking

levels 3

levels 4

levels 2

levels 3

TEST SET SIZE PER CELL – N = 200 TICKETS

every cell classifies its own sample of n tickets: 1 ticket = 1 LLM call

COMPARISONS AVAILABLE

$$11 \times 5 \times 12 = 660$$

EXPECTED NUMBER OF FALSE POSITIVES (THEORY)

$$660 \times 0.05 = 33.0$$

Story Finder

Look at enough numbers and a story will find you. Explore → Claim → Challenge → Compare.

►What is this playground about? (click to expand)

Explore: find stories, then challenge them

Pre-register a comparison instead

The other kind of goal. Exploring asks "what is going on?" and its honest output is hypotheses. Comparing asks "is this condition better than the baseline?", and to make the answer evidence, you fix the question *before* looking: one condition, one metric, one slice, one decision rule, one look. Same simulated reality as the Explore tab.

PRE-REGISTERED COMPARISON

Write down the hypothesis and the decision rule **before** the data exists (uses the conditions, n and reality set on the Explore tab). You get exactly one look. If the target is met, that is evidence; a target chosen afterwards is a description.

Pre-registration

CONDITION (VS BASELINE)

Model A · Prompt variant 2

METRIC

Accuracy

SLICE

All tickets

SHIP IF IMPROVEMENT ≥

2

Decision rule: ship if the improvement is at least the threshold *and* $p < 0.05$ (two-sided). Nothing else counts, however interesting.

Commit & run (one look)

Repeat 20 pre-registered experiments

PRE-REGISTERED RUNS

0

"SHIP" DECISIONS

0

FALSE-SHIP RATE IF NULL

≤ 5%

EXPLORE MODE FOUND

-

Two types of closed-book overfitting

We have two instances of *multiple hypotheses testing* (also called *Closed-book overfitting*):

1. We can “find” stories that do not exist ([story playground](#))
2. **We inflate estimates** ([inflated estimates playground](#))

And: [Playground](#): upload your metrics, see what you can conclude

Inflated Estimates in Multiple Hypotheses Testing

Try N variants on one test set, keep the best. How inflated is its score?

►What is this playground about? (click to expand)

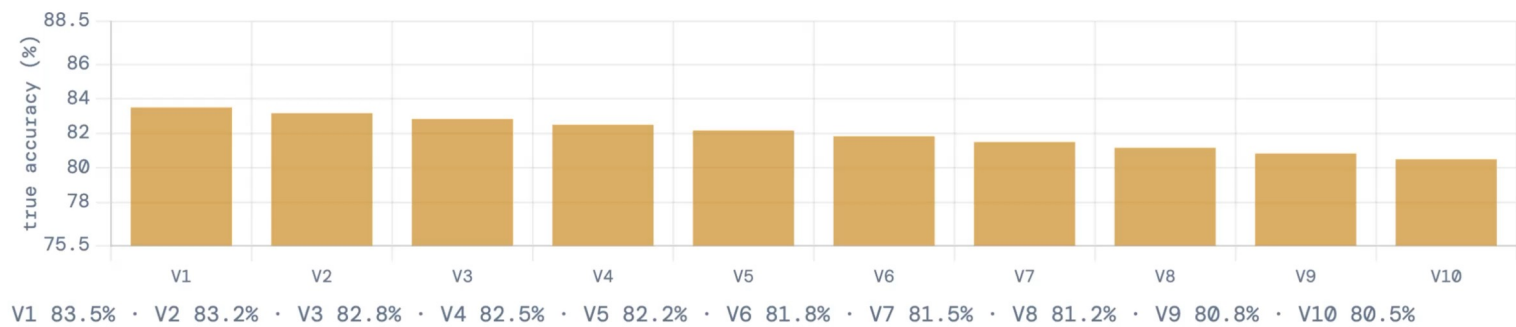
Simulate: pick the best of N, then check it Is my winner really better than the baseline?

Reality: N similar variants, ranked by how good they truly are, based on accuracy.

You have **10 prompt variants** for the ticket router. They are similar, but not identical: **V1 is truly the best** at 83.5%, and they go down evenly to V10 at 80.5%. The differences between neighbours (0.3 pts) are smaller than the noise of a single score (about ± 5.3 points at $n = 200$). You do not know any of this; you only see the scores.

HOW DIFFERENT ARE THEY REALLY? – V1 IS 3 PTS BETTER THAN V10

0 = all exactly equal. Otherwise the true accuracies are spread evenly from V1 (best) down to VN, around 82%.



SETUP (one metric of interest: accuracy of the ticket router)

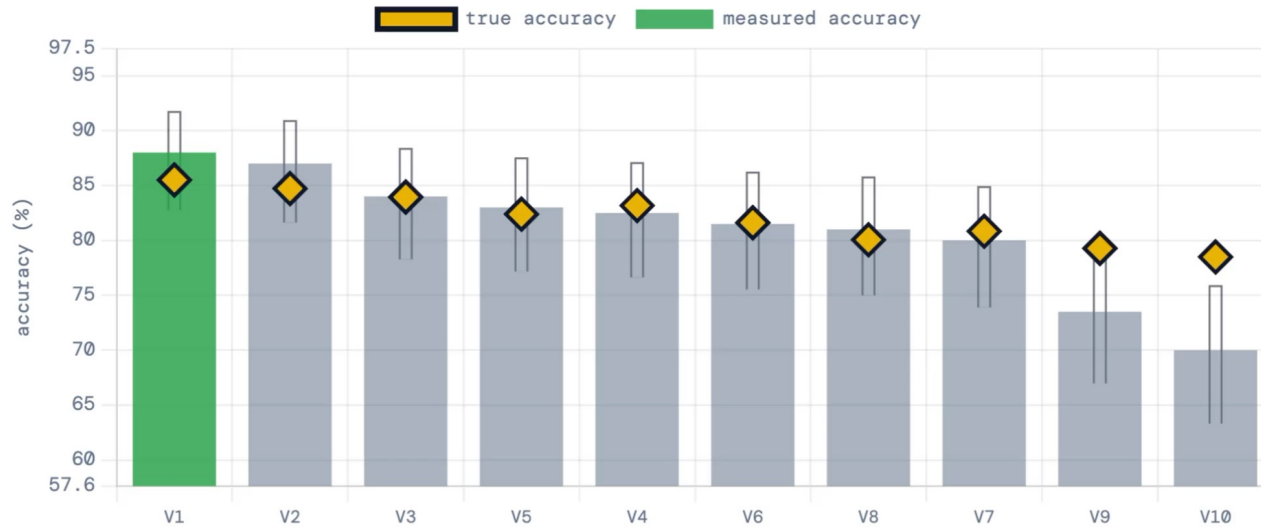
VARIANTS TRIED – N = 10 PROMPT VARIANTS

every variant is scored on the same test set; you keep the one with the highest accuracy

TEST SET SIZE – N = 200 TICKETS

each variant is scored on the same n tickets

1. WHAT WE PICK, AND 2. WHAT WE CLAIM FOR IT



The leaderboard: measured accuracy of each variant (bars, sorted; thin bars: 95% credible interval from a Beta posterior) against its true accuracy (yellow diamonds). Green: the one we pick. Variants are named by their true rank, so V1 is the one we *should* pick.

1. DID WE PICK THE RIGHT VARIANT?

YES we picked V1

V1 is the truly best variant (85.5%).

2. WHAT DO WE CLAIM FOR IT?

85.5%

real accuracy

88.0%

what we estimate

+2.5 pts

overinflation

95% credible interval 82.8% - 91.8% · 176 of 200 right

Outcome: **right pick, overstated**. We ship the best variant, but the number we report is +2.5 pts above what it will deliver. This is the common case: V1 usually wins on the runs where its own luck is good.

RUN 300 EXPERIMENTS

Now, run 300 experiments on 300 different datasets. For each experiment:

1. score all N variants on a *new* test set of n tickets
2. pick the winner
3. estimate its accuracy from that score

▶ Run 300 experiments

Inflated Estimates in Multiple Hypotheses Testing

Try N variants on one test set, keep the best. How inflated is its score?

►What is this playground about? (click to expand)

Simulate: pick the best of N, then check it

Is my winner really better than the baseline?

The question a team actually has. You tried N variants against a baseline, on the same tickets, and one of them came out ahead. Is it really better, given that you tried N? This tab answers with your own data, no assumptions about distributions: it re-samples your tickets thousands of times and asks how big a lead the best of N gets by luck alone.

PASTE YOUR PER-TICKET RESULTS

First row: the names, **baseline first**. Then one row per ticket, one column per system: 1 = right, 0 = wrong (or any score, higher = better). Every system must be scored on the same tickets. Comma, tab or space separated.

No data at hand? Load an example:

Example A: nothing is better

Example B: V3 truly +8 pts

```
baseline, V1, V2, V3
1, 1, 0, 1
0, 1, 1, 0
...
```

WHAT WE CAN SAY WITH "95% PROBABILITY"

Paste results, or load an example, then click Analyze.

Many ways to Impress Your Boss

PICK THE METRIC

Dataset and configuration fixed

Prompt v2 loses on accuracy and on F1. It wins on recall.
The report leads with recall.

In football: pick the stat.

PICK THE CONFIGURATION

Dataset and metric fixed

Nine configurations tie with the baseline. The tenth is up four points.
The tenth ships.

In football: pick the slice.

PICK THE DATASET

Configuration and metric fixed

61% on last quarter's tickets, 72% on the demo set, 84% on the “curated” subset.
The boss sees the 84%.

In football: pick the match.

Regression testing vs Multiple Hypotheses Testing

TRADITIONAL SOFTWARE

Running the same tests on every change is a virtue.

AI SYSTEMS

Running the same test set on every change is how you overfit.

Pre-registration: decide what “good” means before you look

What it is — writing down the goal, the metrics, the thresholds and the decision rule before the results exist, and not changing them afterwards

Why it matters — with 5 metrics, 3 datasets and 4 prompts, something is always good by chance

What it protects — the meaning of the result: a pre-declared target that is met is evidence; a target chosen afterwards is not

How we do it, conceptually: fill the dimensions table (goal, metrics, thresholds, baseline, confidence) and commit it before the first run

Still allowed — exploring the data afterwards is fine; label it as exploration and let it feed the next pre-registered experiment

Repeatability

Record everything needed to run the experiment again and get the same answer: data snapshot, model and version, prompts, parameters, code, evaluation script

If its not repeatable – it is hard to defend it

How we do it, conceptually: one experiment = one frozen bundle (data + config + prompts + code, all versioned and dated) – including the objectives. Ideally, re-run it before reporting to prove it reproduces

Thanks)