DESIGN BEFORE SEARCH

A Hybrid DOE–BO Workflow for LLM Hyperparameter and Prompt Tuning

---

What if the most important tuning decision is knowing what to optimise?

KEY IDEA

- Screen first.
- Optimise second.
- Understand before committing resources.

---

16 DOE Runs

vs.

64 Grid Search Runs

75% fewer

WHAT WE FOUND

- Persona framing mattered most in one task.
- Model choice mattered most in another.
- Different tasks prioritized different factors.

Proof-of-Concept Study

Estimated reading time: 10–12 minutes

AvatarsBio

Design Before Search

A Hybrid DOE–BO Workflow for LLM Hyperparameter and Prompt Tuning

Chee Ping Ng, Founder
AvatarsBio, The Netherlands

LLM-based applications expose many tuneable factors at once — prompt structure, persona framing, instruction style, output format, temperature, model choice — and exhaustively tuning across all of them grows expensive and increasingly difficult to interpret as the factor count increases. This white paper presents a hybrid workflow in which Design of Experiments (DOE) screens a structured, balanced subset of configurations to identify which factors drive performance, and Bayesian optimisation (BO) is applied afterward to refine within that narrowed space. We demonstrate this workflow on two tasks: sentiment classification (SST-2), where a contamination-driven ceiling effect showed the dataset, not the method, was the limiting factor; and molecular permeability classification from chemical structure (Caco-2), where screening found a real but modest signal and showed that the most influential tested factor changed when the task changed. The contribution is the workflow rather than a new algorithm, and not a claim about LLM suitability for either task. A structured screen, run before optimisation begins, identifies what is worth optimising and provides more insight than exhaustive tuning alone, at a fraction of the run cost.

THE TUNING PROBLEM

LLM-based applications expose many tuneable factors at once — prompt structure, persona framing, instruction style, output format, temperature, model choice — each plausibly affecting how well a task is performed. The common response is grid search: run every combination, keep the best score.

Grid search cost grows multiplicatively with the number of factors. Six binary factors already mean 64 runs; a seventh doubles that again. Cost aside, grid search returns a leaderboard, not an explanation. It tells you which configuration scored highest, not which factor was responsible, whether the runner-up's lower score is a real difference or noise, or whether the ranking would survive changing one factor while holding the others fixed. A practitioner inheriting that leaderboard later must either trust it blindly or reconstruct the reasoning from scratch.

Design of Experiments (DOE) starts from a different question: what is the smallest structured set of runs that can estimate each factor's individual contribution [1], before committing to an exhaustive sweep? This logic is not new to this domain — it extends a philosophy already applied in biological modelling, set out in a related white paper, Design Before Prediction: reduce uncertainty through structured screening before committing resources to optimisation [2]. The remainder of this paper applies the same logic to LLM tuning: DOE narrows the field of plausible factors first; Bayesian optimisation (BO) then refines within the narrowed space [3], on the continuous dimensions DOE can only coarsely sample.

DESIGN BEFORE TUNING

Design Before Prediction argued that structured screening can reduce uncertainty before resources are committed to model development. The same logic applies here. Rather than

treating optimisation as the first step, DOE is used to identify influential factors before Bayesian optimisation begins (Figure 1).

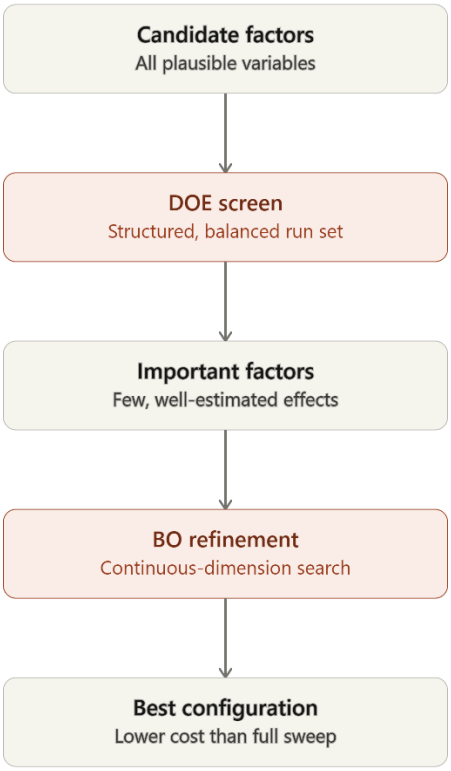

Figure 1. Hybrid DOE–BO workflow. A structured screening phase uses DOE to identify influential factors from a large candidate set. Bayesian optimization is then applied only to the reduced factor space, focusing optimization effort where it is most likely to be useful.

DOE is not a post-hoc analysis step run after the real work is done. As in Design Before Prediction, it is part of planning the work itself — applied here to LLM tuning factors rather than biological variables [2].

DOE AS A SCREENING LENS FOR LLM FACTORS

Design of Experiments is a structured way to plan a set of runs so that each factor's effect can be estimated without testing every combination [1]. Classical screening designs — fractional factorials, Plackett-Burman designs — were developed for exactly this problem: many candidate factors, a limited run budget, and a need to separate signal from noise before committing to a full sweep [4].

The same logic applies to LLM tuning, with a different vocabulary. Where a biological screen might vary matrix stiffness or growth factor concentration, an LLM tuning screen varies factors like:

- Prompt style — neutral instruction vs. a stated persona or role
- Instruction framing — direct request vs. chain-of-thought reasoning
- Example inclusion — zero-shot vs. one worked example
- Output format — free text vs. a strict required format
- Temperature — sampling randomness
- Model — which underlying model serves the request

Not all factors matter equally, and which ones matter is rarely obvious in advance. A factor that dominates on one task may be negligible on another; an interaction between two factors may matter more than either factor alone. Treating tuning as a screening problem — run a structured, balanced subset of configurations, estimate each factor's effect, then decide what is worth optimising further — replaces guesswork with a documented answer to a simple question: which factors are driving performance?

This is the same shift DOE makes in experimental biology: moving from "test everything" or "trust intuition" to a smaller, structured set of runs that tells you where to look next.

DEMONSTRATION

The workflow was first applied to SST-2 sentiment classification, mainly to confirm the DOE→BO pipeline ran correctly end to end. An initial screen on 50 examples found accuracy ranging from 0.94 to 0.98 across configurations. A follow-up run pushed every configuration to near-ceiling performance, leaving no meaningful separation between factors — consistent with contamination of this widely benchmarked dataset. The lesson: a benchmark a model can already solve gives a screening method nothing to screen. The pipeline was validated; the task was not.

The main demonstration is a harder and more practically relevant task: predicting Caco-2 permeability directly from a SMILES string, with no chemistry-specific tooling. Unlike SST-2, there is little reason to expect the model has memorised the task, and the underlying biology gives the result relevance beyond a pure methods demonstration.

A first screen, varying four prompt-level factors across all 16 combinations, found that giving the model a stated professional identity ('you are a medicinal chemist...') was the dominant factor (Figure 2), improving accuracy by roughly 9 percentage points on average, well ahead of reasoning style or worked examples, both of which had little effect on their own. Forcing a strict answer format was actively harmful, often cutting the model off before it reached a conclusion. The best configuration reached a balanced accuracy of 0.633 against a 0.5 chance baseline: a real but modest signal, consistent with asking a language model to do something closer to pattern-matching on chemical structure than physics-based prediction. While the resulting classifier is not competitive with dedicated QSAR models, its utility may lie in early-stage planning and triage of clearly high- or low-permeability compounds rather than resolving ambiguous cases.

That factor importance can shift with task context is unsurprising in principle. The question is whether a practitioner knows which factor will matter before checking. Extending the task to a three-class version (low/mid/high permeability), the screen was widened to six factors — the original four prompt-level factors, plus temperature and, for the first time, choice of underlying model — using a true fractional design: 16 runs standing in for the 64 a full sweep would require. The result that matters most for this paper was that the most influential tested factor changed. Persona framing was the strongest factor among those screened in the binary task; once model choice was added to the screen, it became the strongest effect, while persona's own effect fell by more than half (Figure 2) without disappearing. Model wasn't a previously-tested factor that lost ground — it was a factor nobody had screened for at all in the binary task. A team that had tuned around persona framing because it worked once would have had no structured reason to add model choice to its short list when the task changed; the wider screen caught it directly, without relying on anyone noticing.

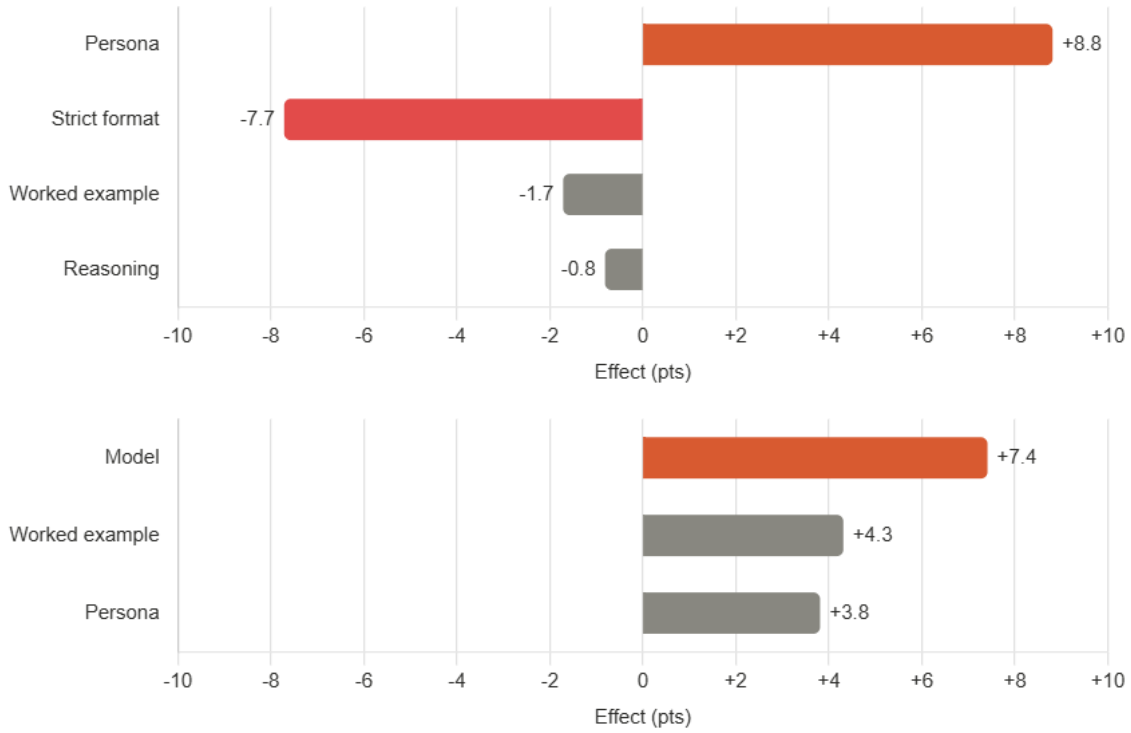

Figure 2. Dominant factors changed when task difficulty changed. In the binary permeability task (top), persona framing produced the largest measurable effect on performance. In the three-class task (bottom), model choice became the dominant factor. The result illustrates that different tasks can elevate different factors, reinforcing the value of structured screening before optimisation.

Bayesian optimisation was then applied to the factors the screen had identified as worth pursuing further — model, example inclusion, prompt style, and temperature — not to the full original factor list. It found a modest further improvement (0.592, up from the screen's best of 0.567) and converged within its run budget, though the optimal temperature landed at the edge of the range searched, leaving open whether a wider range would have done better. BO's role here is refinement of an already-narrowed space, not the source of the paper's main finding.

WHAT THE HYBRID WORKFLOW ADDS

Two benefits extend beyond the specific demonstration and carry over to any LLM tuning project of comparable scale.

The first is a structural reduction in wasted runs, independent of which configuration happens to win. A full sweep across six binary factors is 64 runs; the fractional screen used here reached a usable answer — which factors mattered, and roughly by how much — in 16 (Figure 3). More importantly, it reduced the risk of tuning around the wrong factor. Without a structured screen, a practitioner might reasonably optimise around whichever factor appears responsible for an early result, only to discover later that a different factor was driving performance. Screening protects against that kind of misattribution by estimating each factor's effect directly rather than simply reporting which configuration scored best. The honest caveat: this reduces rework on the question of which factors matter, not total project effort. Confound-checking, pilot diagnosis, and interpretation still take time, with or without DOE.

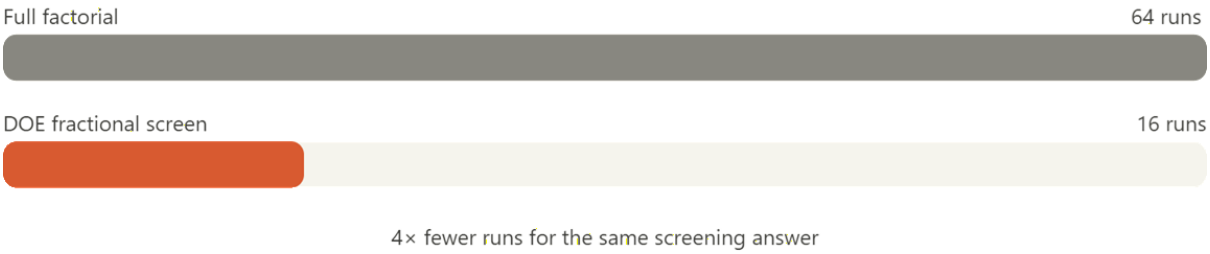

Figure 3. Exhaustive tuning versus structured screening. A full sweep of six binary factors requires 64 configurations. The fractional screening design used here estimated factor effects using only 16 runs before optimization began.

The second benefit is less obvious, and arguably more durable: a documented reason for the result, not just the result. A grid search produces a leaderboard — the best score, and the runs that didn't make the cut. It doesn't explain why the winner won, whether the runner-up was a real contender or noise, or whether the ranking would survive changing one factor at a time. Anyone picking the work back up later has to either trust the leaderboard or re-derive the reasoning from nothing. The DOE structure used here instead leaves a record: which factors mattered and by how much, which interactions couldn't be separated and were flagged rather than ignored, and where a pipeline quirk — a model that stopped predicting one class entirely under one framing — was a real finding rather than noise to be cleaned up. That record is the difference between a result and something a collaborator can actually audit.

LIMITATIONS

This study is not intended as a rigorous benchmark. Sample sizes throughout were small — 50 SST-2 examples, 44–90 Caco-2 compounds — and the permeability task ran without replicate runs at each design point, so the effect sizes reported here should be read as indicative, not definitive.

The screening designs used resolution IV fractional factorials in two of the study's stages. At this resolution, several two-factor interactions are statistically indistinguishable from each other and from aliased three-factor interactions — a deliberate trade-off for run-count efficiency, but one that means some "this factor matters" conclusions could in principle be confounded with an interaction rather than reflecting a clean main effect.

The three-class permeability labels also carry a known confound: the physicochemical descriptors that separate low/mid/high classes (LogP, TPSA, hydrogen-bond counts) are the same kind of information a language model could plausibly be drawing on when predicting class membership from a SMILES string. This doesn't invalidate the screening result, but it does mean the model and the labeling scheme may be drawing on related information.

One claim made informally during the binary Caco-2 work — that a half-fraction design would have found the same key factors at half the run cost — was inferred after the fact from the full 16-run result, not confirmed by independently re-analyzing a held-out 8-run subset. It's a reasonable expectation, not a demonstrated one, and is presented that way.

Taken together, this study is best read as a demonstration of the DOE→BO workflow across a small number of tasks and model families — not as a benchmark conclusion about LLM suitability for molecular permeability prediction, or about which prompt factors matter universally for LLM tuning.

CONCLUSION

DOE before optimisation is a workable structure for LLM hyperparameter and prompt tuning. The screening designs used here estimated which factors mattered with a fraction of a full grid sweep's

run count, surfaced a finding — different tasks elevating different factors — that exhaustive tuning would not have made visible on its own, and left behind a documented reason for each result rather than just a ranked list. Bayesian optimisation, applied afterward to the factors screening identified as worth pursuing, added a further but modest refinement.

What this study does not show is that LLMs are well-suited to molecular permeability prediction, or that any one factor — persona framing, model choice — will dominate on tasks beyond the ones tested here. Those are separate claims, and the small samples and single-run designs used wouldn't support them regardless. A fuller technical treatment — complete factor tables, confidence intervals, formal interaction analysis — is a natural next step, distinct from what this paper sets out to do.

As AI systems expose an increasing number of tuneable parameters, screening may matter not only for finding better configurations, but for identifying which factors are worth optimising at all.

> ***Understanding a system may be as valuable as improving it.***

The contribution here is narrower than either claim, and still useful: showing that a structured screen, run before optimisation starts, tells a practitioner something exhaustive tuning alone does not.

> ***Screen first. Optimise second. Trust the numbers least of all.***

SELECTED REFERENCES

1. Montgomery, D. C. (2019). Design and Analysis of Experiments (10th ed.). Wiley.

2. Ng, C. P. Design Before Prediction: A hybrid DOE-ML workflow for reducing complexity in biological models. AvatarsBio white paper. https://avatarsbio.com/research/pdfs/design-before-prediction-doe-ml.pdf

3. Frazier, P. I. (2018). A tutorial on Bayesian optimization. arXiv:1807.02811. https://arxiv.org/abs/1807.02811

4. Plackett, R. L., & Burman, J. P. (1946). The design of optimum multifactorial experiments. Biometrika, 33(4), 305–325. https://doi.org/10.1093/biomet/33.4.305