PG-Agent: can frontier LLMs order protein variants by fitness?

ABSTRACT

Frontier large language models (LLMs) are increasingly proposed as zero-shot predictors of protein-variant fitness, yet the claim has never been tested under controlled, baseline-matched conditions. We introduce **PG-Agent**, a contamination-audited benchmark that asks whether an LLM can rank deep-mutational-scanning variants from sequence alone, with no alignments, structures, fitness labels, or mutation annotations. Across 217 ProteinGym assays, ten frontier LLMs and 95 specialized predictors are scored on byte-identical, label-stripped subsamples. The LLMs are capable but outclassed: the best reaches only 64% of the leading baseline's correlation (nested-macro Spearman $\rho = 0.34$ versus 0.53), sits at the 18th percentile of predictors, adds almost nothing in ensemble, and decays further as the candidate set grows. We trace the ceiling to mechanism: about half of an LLM's ranking is captured by a six-feature substitution-severity prior, so its confident structural narratives substantially reflect a simple heuristic. Across 1,423 reasoning traces that narrative neither improves accuracy nor, despite frequent recognition of the source experiment, corrupts the scores. Frontier LLMs are fluent narrators of variant biology and unreliable rankers of it.

1 Introduction

Directed evolution and rational design reduce to one question: among many candidate sequences, which are most fit? Specialized predictors answer it well, and ProteinGym[1] is the field's standard yardstick for alignment-based models (EVE[2], GEMME[3]), protein language models (ESM-1v[4], ESM-2[5]), and structure- or retrieval-conditioned models (SaProt[7], VenusREM[9]). General-purpose LLMs are now proposed for the same role, and the PG-Hard evaluation in a recent system card[10] reported that a frontier model can rank a handful of variants above chance. Whether that ability is real, useful, or robust has not been measured against baselines under controlled conditions.

We introduce **PG-Agent**, a benchmark that gives a model only what it would have on a novel design problem: the wild-type and N full mutant sequences, with no alignments, structures, fitness labels, or mutation annotations. Every model and 95 published predictors are scored on byte-identical, label-stripped subsamples across candidate-set sizes from 10 to 500. Beyond accuracy, we ask *how* models rank, open-coding 1,423 reasoning traces and auditing them for memorized knowledge of the specific experiment. Our contributions are:

- **PG-Agent**, a controlled, baseline-matched, sequence-only benchmark for LLM variant ranking with a 10-to-500 candidate-set scaling axis.
- **A mechanistic probe** showing that about half of an LLM's ranking is captured by six trivial substitution features, dominated by conservative-substitution score.
- **A reasoning audit** that links strategy to accuracy (loss-of-function inversion is the worst), and finds widespread dataset recognition but no detectable score-corrupting leakage.

2 Related work

Variant-effect prediction. Deep mutational scanning[11] measures thousands of variants per assay, and ProteinGym[1] aggregates 217 substitution assays scored by a nested-macro Spearman that equal-weights functional categories and de-duplicates multi-assay proteins. One line of predictors is alignment-based, including EVmutation[12], DeepSequence[13], EVE[2], GEMME[3], and the MSA Transformer[14]. A second line uses protein language models, including ESM-1v[4], ESM-2[5], Tranception[15], ProGen[16], and CARP[17]. A third conditions on structure or retrieval, including ESM-IF1[6], SaProt[7], TranceptEVE[8], ProSST[18], and VenusREM[9], the strongest predictor in our re-scoring. All of them receive inputs an LLM does not: explicit alignments, structures, or per-variant scoring.

LLMs as scientific reasoners and rankers. General LLMs are increasingly applied to biological reasoning and design ideation, and the PG-Hard evaluation in a recent system card[10] reported above-chance variant ranking by a frontier model; PG-Agent generalizes it into a controlled, baseline-matched, multi-model benchmark with explicit scaling. The task also connects to work on LLMs as listwise rankers, which documents position bias and a lost-in-the-middle degradation as the list grows[19,20], and to the chain-of-thought faithfulness literature, which shows that a model's stated reasoning is often not the cause of its answer[21,22]. Our mechanistic probe and recognition audit give that phenomenon a sharp, domain-grounded instance.

3 The PG-Agent benchmark

3.1 Task

For a DMS assay with wild-type sequence and a set of substitution variants, we draw a subsample of N variants and present the model with the wild-type sequence and the N full mutant sequences in shuffled order, each given an arbitrary integer label (M01 to M50 at $N = 50$). The prompt states the protein name, organism, what the assay measures, and the convention that higher fitness means a higher value of that property; it includes no fitness values and, by default, no mutation short-

hand, so the model must diff the sequences itself. The model reasons and then emits, on the final line, a JSON ranking, which we score by Spearman ρ against the held-out DMS fitness.

3.2 Data and stratified subsampling

We use the full ProteinGym v1.3 substitution set: 217 assays over 186 proteins, spanning five functional categories (Organismal Fitness, Stability, Activity, Expression, Binding), four taxa, and both single- and multi-mutant assays. For each assay and candidate-set size N (10, 50, 100, 500) we draw a fitness-stratified subsample of N variants with a fixed random seed, then strip the labels. The fixed seed means every model and every baseline sees exactly the same variants. At N of 10, 50, and 100 we draw three independent subsamples per assay, and the spread between them gives the reproducibility estimate reported in §4.9.

3.3 Metric, models, and baselines

The headline metric is ProteinGym's nested-macro Spearman ρ. To keep any single large assay or frequently-studied protein from dominating, we average in stages rather than pooling all variants: per-assay ρ, then the mean within each UniProt protein, then the mean within each of the five functional categories, then the unweighted mean of those five. Spearman ρ ranges from -1 to 1, where 0 is chance; the best predictor here reaches ∼0.5. The headline ρ collapses the three runs to one value per assay before aggregation, and error bars are the SEM across the three runs unless noted. We evaluate ten frontier LLMs through native provider APIs at high reasoning effort (GPT-5.5, 5.4-mini, 5.4-nano; Claude Opus 4.8, Opus 4.7, Sonnet 4.6; Gemini 3.5 Flash, 3.1 Pro, 3.1 Flash-Lite; and GLM-5.2), plus an effort sweep on GPT-5.5, Gemini 3.5 Flash, and Claude Opus 4.8, and an N = 500 sub-study on the Gemini family. We re-score 95 ProteinGym baselines on the identical frozen subsamples. All models run at temperature 1 at each provider's high reasoning-effort tier for the headline numbers; we parse the ranking from the JSON on the final line and count refused or unparseable responses as missing. Per-cell coverage is reported with every result; API refusals of viral or pathogen sequences (GPT models) and empty responses (Claude Sonnet at large N) reduce some cells below full coverage, which we flag rather than hide (Table S1). For the reasoning analyses (§4.7, §4.8) we additionally captured the models' reasoning channels in a dedicated N = 50 re-run (OpenAI reasoning summaries, Gemini thought-summaries, and Claude summarized thinking), since the main runs persisted only final answers.

4 Results

4.1 Frontier LLMs are capable but far from specialist-level

Figure 1 ranks the ten LLMs at N = 50 against the 95 specialized baselines on identical subsamples. Gemini 3.5 Flash tops the full-coverage ranking at ρ = 0.34, but the lead is within the benchmark's resolution: a paired bootstrap over assays cannot separate the top five (Gemini 3.5 Flash, Gemini 3.1 Pro, GPT-5.5, Claude Opus 4.8, and Sonnet 4.6, all ρ ≈ 0.31 to 0.34), and on the 131 assays every model covers, GPT-5.5 and Gemini 3.5 Flash are a dead heat. We therefore read the top as a single tier rather than a strict order; below it the field falls more than twofold to GPT-5.4-nano (0.15) (Table S1). In absolute terms this is a real capability: with no MSA, structure, or labels, a frontier model orders variants at ρ = 0.34, well above chance and above 17 of the 95 specialized predictors. Those 17, however, are the field's weakest, the small or retrieval-free protein language models. In relative terms the best LLM beats only 17 of the 95 baselines (the 18th percentile, below the first quartile of 0.36): more than four-fifths of specialized predictors do better, and it reaches only 64% of VenusREM's correlation (0.527). Two caveats keep the comparison fair. First, the baselines use information the LLM is denied (alignments, structures, or retrieval) and score each variant independently, whereas the LLM must commit to one joint ordering (§4.2). Second, coverage differs across models: GPT-5.5 scores 199 of 217 assays (refusing 18 viral or pathogen sequences) and Sonnet 4.6 scores 135 (empty responses at large N; Table S1), so the ordering is indicative rather than exact, especially for the closely-spaced top tier. The honest summary is capable in isolation, dominated in context.

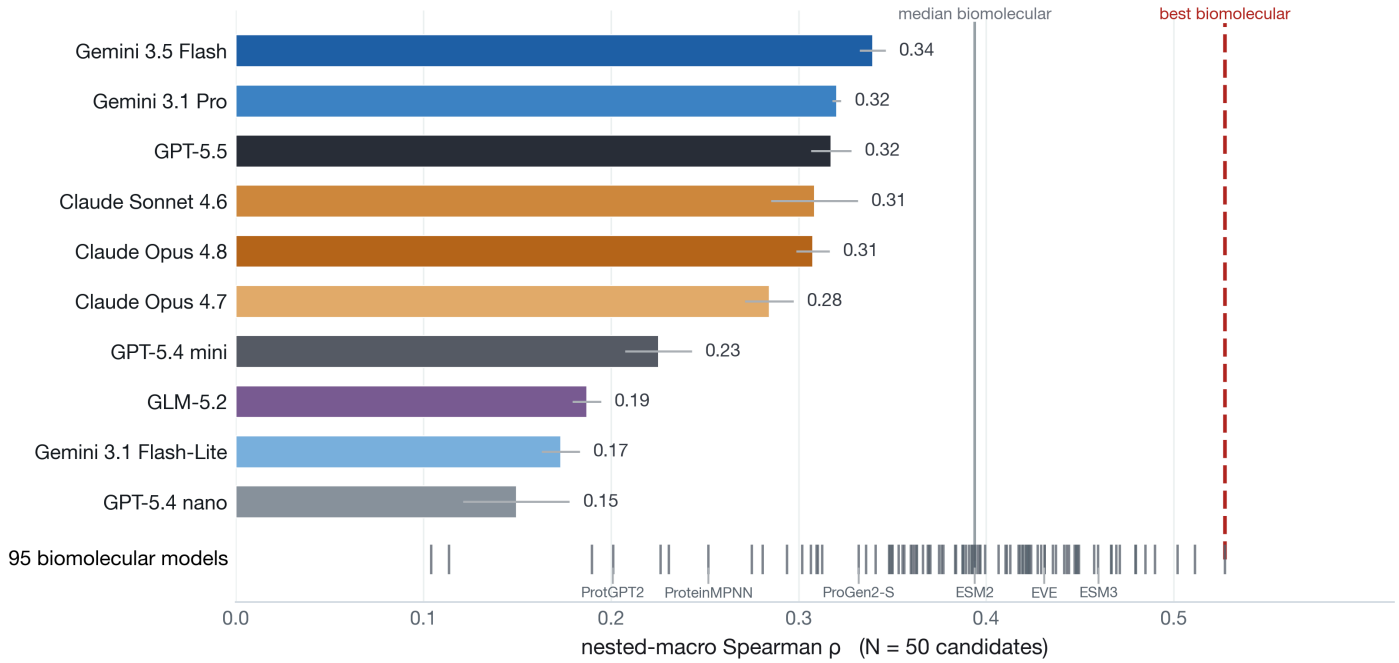

Figure 1 | Leaderboard at N = 50. Nested-macro Spearman ρ for the ten LLMs (error bars: SEM across the three runs), coloured by provider with a shade per model. Ticks below the axis show all 95 biomolecular models; the grey line marks their median (median biomolecular, 0.39) and the dashed crimson line the best (best biomolecular, VenusREM, 0.53), both recomputed on the same subsamples. Named for reference: models the LLMs now match or beat (ProtGPT2, ProteinMPNN, ProGen2-S) and bigger ones not yet surpassed (ESM2, EVE, ESM3). Every LLM falls below the median: the best beats only 17 of 95 predictors (18th percentile), lands on par with small protein language models, and reaches 64% of the best model's correlation.

Per-size, single/multi, and SEM breakdowns are in Table S1; single- and multi-mutant accuracy are similar for most models (within the sampling spread), with no clean provider pattern, so we do not read the small differences as evidence about epistasis.

4.2 Ranking accuracy collapses with set size

The most diagnostic axis is set size, and it reflects a structural asymmetry between the two model classes: a specialized predictor assigns each variant an independent score, so set size is irrelevant to it, whereas an LLM asked for a single ordering must hold all N candidates in context and emit one joint permutation, a fundamentally harder, capacity-bound task. The consequence is stark (Figure S1). Baselines are flat in N, while every LLM degrades from N = 10 to 100 (Table S1); the three Gemini models we extend to N = 500 fall further still, Gemini 3.5 Flash from 0.395 to 0.188 and Gemini 3.1 Flash-Lite from 0.292 to 0.019, indistinguishable from random (Fig S1). The collapse is capacity-dependent, with the weakest model losing the largest fraction of its small-N skill, which is consistent with the joint-permutation demand rather than per-variant judgement as the bottleneck, though we did not test this against pointwise elicitation (§6). We measure this under joint-ranking elicitation, the natural way to ask an LLM for an ordering; whether eliciting an independent score per variant (pointwise or pairwise) would recover the signal is a control we did not run (§6). Under joint ranking, this is the cleanest separator between general-purpose rankers and specialized scorers, and it is bad news for design, where candidate pools are large.

4.3 Reasoning helps with steep diminishing returns, at very different token costs

Spending more reasoning improves ranking, but with sharply diminishing returns and very different efficiency across models (Figure 2). Sweeping the effort tiers of GPT-5.5, Gemini 3.5 Flash, and Claude Opus 4.8 and logging the output tokens each spends (including hidden reasoning), every model gains most of its skill in the first few thousand tokens and then flattens: GPT-5.5 rises from ρ = 0.24 to 0.32 and Gemini from 0.25 to 0.34. The sharper result is token efficiency. Claude Opus 4.8 reaches ρ = 0.31 at just 4.9k tokens (its medium tier), matching GPT-5.5's high tier (0.31 at 18.7k) at roughly a quarter of the tokens, and tops out near ρ = 0.34 with 16.9k tokens, with its maximum tier spending 2.2 times as many tokens (37.8k) for no further gain (0.34); Gemini is the opposite, spending 28.7k tokens to reach the same level. Tier-to-tier accuracy differences are within the benchmark's resolution, so the figure is best read as token spend at matched accuracy rather than an accuracy ranking; the sweep is a separate sub-study whose absolute ρ values are not directly comparable to the N = 50 leaderboard. We report tokens, not dollars, since per-token prices differ by provider and change often. Within the tested range (up to about 38k tokens, Claude Opus 4.8's maximum tier) more reasoning did not close the gap to VenusREM, so the ceiling is not simply a matter of test-time compute, though we cannot exclude that decomposition-based methods would help.

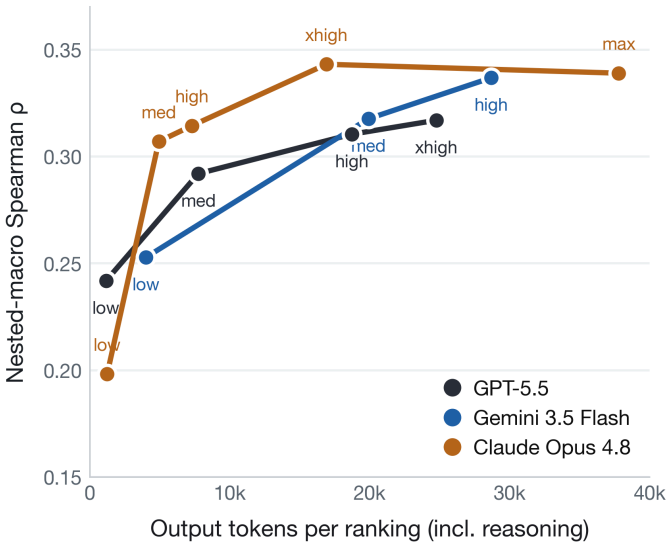

Figure 2 | Reasoning effort buys little accuracy, at very different token costs. Nested-macro ρ against mean output tokens per ranking (including hidden reasoning) at N = 50, sweeping each model's effort tiers (low to maximum; Gemini exposes no tier above high). Accuracy saturates within a few thousand tokens for every model, but the token cost of a given accuracy differs several-fold: Claude Opus 4.8 is the most token-efficient and Gemini 3.5 Flash the least. Opus 4.8's maximum tier spends 2.2 times the tokens of its extra-high tier (37.8k vs 16.9k) for no gain, so the plateau is not a budget artifact. Tier-to-tier differences are within the benchmark's resolution; this effort sweep is a separate sub-study not directly comparable to the N = 50 leaderboard.

4.4 LLMs do not complement the specialized field

If LLMs solved different assays than specialized models, they would be useful even while weaker on average. They do not (Figure 3). The same assays are hard for both: per-assay, Gemini 3.5 Flash and VenusREM are correlated at r = 0.58, and the LLM sits below the diagonal almost everywhere. It beats VenusREM on only 9% of assays and the per-assay best of the 95 baselines on under 1%. Even an oracle ensemble that routes each assay to whichever of VenusREM and the LLM is better, an upper bound on complementarity, improves on VenusREM alone by at most 0.005 ρ for any LLM. The LLM is a dominated signal, not a complementary one, so the common proposal to use it as a cheap pre-filter or ensemble member adds cost without information.

Figure 3 | LLMs do not complement the specialized field. Per-assay ρ of Gemini 3.5 Flash versus VenusREM (N = 50, colored by function). The two are correlated (r = 0.58): the same assays are hard for both, and Gemini lands below the diagonal almost everywhere, beating VenusREM on 9% of assays and the per-assay best of 95 baselines on under 1%.

4.5 What LLMs are doing: a substitution-severity heuristic

To see through the prose, we decoded every variant into six elementary substitution features (BLOSUM62 score, absolute changes in hydropathy, volume, and charge, relative position, and terminus distance) and asked how much of each model's ranking they explain. The LLM's ordering correlates with BLOSUM62 at r = 0.47, well above how well BLOSUM62 itself predicts fitness on these assays (r = 0.17). The fuller picture comes from an OLS fit on all six features: it captures 0.44 to 0.51 of the LLM's ranking variance across Gemini, GPT, and Claude. (The same features explain only 0.24 of the noisier experimental fitness, but that comparison partly reflects how much smoother a deterministic LLM ordering is than a noisy assay; the load-bearing point is the absolute level, that roughly half of an LLM's ranking is six trivial features.) That ranking is therefore largely a substitution-severity prior, with conservative-substitution score its largest single term, applied more uniformly than fitness behaves. This reconciles the respectable floor (severity is a decent prior) with the firm ceiling (it is a poor discriminator).

4.6 Where it breaks: the high-fitness regime and category structure

Category structure is consistent (Table S2): Stability and Expression are most tractable, while Binding and Organismal Fitness, where fitness is an indirect readout, are hardest, and the ranking of models is broadly preserved within each category. The sharper failure appears off the uniform grid. When we re-draw the candidate set toward the high-fitness tail (a top-5%-or-50 pool, closer to a late-stage design campaign), Gemini 3.5 Flash falls to ρ = -0.04, indistinguishable from zero given the sampling spread, with 62% of assays below zero. This is not specific to LLMs: VenusREM collapses identically, from 0.52 to 0.04, the baseline median falls to 0.00, and 46 of 95 baselines go negative. Once the pool is compressed to high-fitness vari-

ants, the remaining differences are within the benchmark's resolution, and no current method ranks them reliably. Part of this is mechanical: restricting to high-fitness variants shrinks the score range, which deflates Spearman ρ for any method, so the honest reading is that little rankable signal survives this regime, not that every method actively fails. Either way, the regime where engineering operates is essentially unsolved, and LLMs do not change that.

4.7 How models reason: emergent strategies linked to success

To characterize how models reason, we ran a bottom-up, two-stage pipeline over one full N = 50 run per reasoning-exposing model (1,423 traces, seven models, four providers, with GPT-5.5 included via its reasoning summaries). Stage one open-codes each trace into a free-text description of its ordering logic, with no preset categories; stage two groups those descriptions into the strategies that recur, assigning each trace a single label. Both stages use an LLM judge (Gemini 3.5 Flash) given only the trace text, blind to the assay's labels and the trace's ρ. Nine substantive strategies emerge and differ sharply in how well they rank (Figure 4a); a tenth group is bare numeric scoring, a presentation style we omit, and a small remainder (together about 6% of traces) are unclassified. These links are correlational: stronger models favour particular strategies, and, as §4.5 shows, a stated strategy need not be what actually drives the ranking. We therefore read Figure 4a as which approaches accompany better rankings, not as proof that adopting them causes better rankings.

The lowest-scoring strategy is getting assay polarity backward. Loss-of-function inversion is the only labelled strategy that scores below chance (mean ρ = -0.12, n = 67 traces, single-judge labels; §6): when a model recognizes an inverted assay (a loss-of-function selection such as p53 or CcdB) and deliberately flips its damage-to-fitness logic, it tends to flip it the wrong way. This is the high-fitness failure of §4.6 seen at the level of individual traces.

Generic strategies accompany lower accuracy than concrete ones. The two lowest-scoring positive strategies are the pure substitution-severity heuristic (ρ = 0.21) and opaque bare lists with no stated rule (ρ = 0.22), matching the §4.5 finding that a BLOSUM-like prior is a poor discriminator. The strategies that accompany the best rankings instead commit to something concrete and assay-appropriate: mutation-count priority on multi-mutant sets, transmembrane-hydrophobicity reasoning, and helix biophysics (ρ ≈ 0.33 to 0.36, Figure 4a).

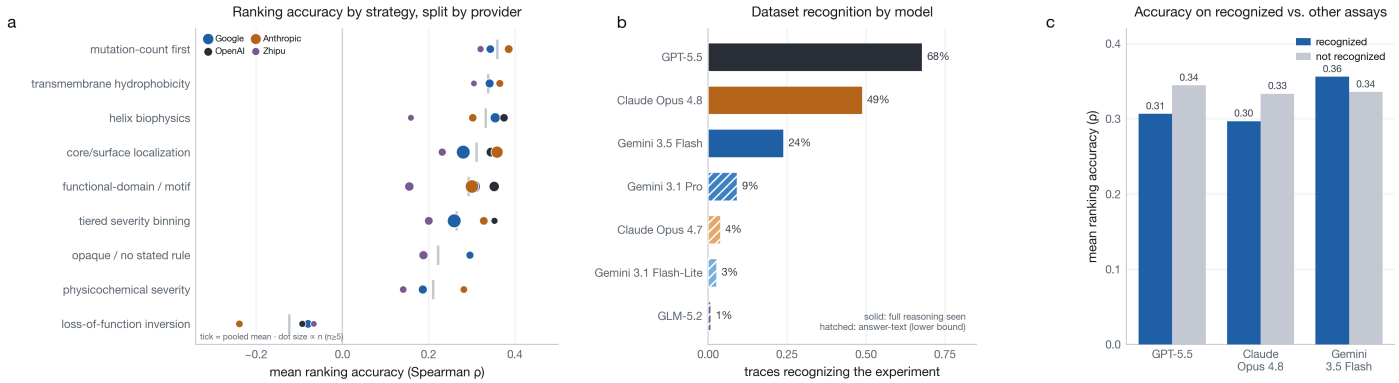

Figure 4 | How models reason: strategies that work, and recognition without corruption. Bottom-up analysis of 1,423 reasoning traces (seven models, four providers; GPT-5.5 via its captured reasoning summaries), each open-coded into a free-text description and grouped into emergent strategies with no preset categories, one label per trace (the nine substantive strategies are shown; a bare numeric-scoring group and a small unclassified remainder, together about 6% of traces, are omitted). **(a)** Mean ranking accuracy (Spearman ρ) per strategy, split by provider (one dot per provider, sized by trace count, shown where n ≥ 5; grey tick = pooled mean): **loss-of-function inversion is the only strategy below chance** across providers, and within most strategies GLM-5.2 (Zhipu) is the weakest provider, so the pooled per-strategy ranking partly reflects which models favour each strategy. **(b)** Dataset-recognition rate per model (dark = full reasoning channel captured; light = answer-text only, a lower bound). **(c)** Mean ranking accuracy on the assays each model recognizes versus those it does not: recognition gives no lift (Δρ ≈ 0, slightly negative for GPT-5.5 and Opus 4.8), so it does not corrupt scores.

4.8 Recognition versus leakage: contamination without corruption

Because the prompt names the protein and assay, a model could recall the specific experiment rather than reason about it. We separate two events. **Dataset recognition** is the model naming or recognizing the specific study, dataset, or benchmark, a contamination of its knowledge; **value leakage** is the stronger event where recalled per-variant values actually drive and inflate the ranking, a contamination of the scores.

Recognition is common where reasoning is visible. Each recognition flag required a verbatim quote; for 89% we confirmed the quote appears word-for-word in the model's own reasoning, citing facts it was never given (such as "a Rocklin 2017 study," "datasets from Giacomelli et al. 2018 in ProteinGym," and "positions 39, 40, 41, 54, as in Wu et al., 2016"). Measured ungated over all 1,423 traces, recognition reaches 68% of GPT-5.5 traces, 49% for Claude Opus 4.8, and 24% for Gemini 3.5 Flash (Figure 4b). For the remaining models we see only terser answer-text, so their rates (1% to 9%) are lower bounds, and cross-model comparison is confounded by reasoning-channel visibility rather than by behaviour alone.

Recognition does not become detectable leakage. If a model were recalling the true fitness values, it would beat the special-

ized baselines on the assays it recognizes. None do: no trace beats the best of 95 baselines by more than +0.19 ρ on any single assay, and that lone extremum is not systematic. The within-model test points the same way: the assays a model recognizes score no higher than the ones it does not (Figure 4c; Δρ = -0.04 to +0.02, and every model's 95% bootstrap interval includes zero), though recognized assays are partly confounded with intrinsic difficulty, so we read this as consistent with, not proof of, no leakage. Even when a model lists specific variants it claims to remember, the recalled effects are too few and too noisy to move a 50-way ranking, so we can exclude leakage large enough to dominate a ranking, while remaining unable to detect silent recall below roughly 0.1 ρ (about a third of the best LLM's total signal of 0.34); leaderboard-magnitude silent recall is therefore bounded, not excluded. For a benchmark built on a published corpus this is the reassuring result: models often know which experiment they are looking at, but that knowledge does not inflate their scores. (We detect stated recall, not silent use; GPT's visible text is a provider summary, which likely raises its recognition rate relative to the others.)

4.9 Benchmark validity

Three threats could distort these results, and each is bounded. **Sampling spread.** A single assay's ρ wobbles by about 0.10 depending on which variants are drawn (this is subsampling variance, distinct from the assay's own measurement noise), but that variation is independent across assays and seeds, so the nested-macro headline is stable to 0.015 across the three seeds, which sets the minimum detectable difference between adjacent models at roughly 0.03 to 0.05. The error bars we plot are the SEM across those three runs; the cross-protein generalization error (a category-level bootstrap, with only five categories at the top level) would be wider still, so we read adjacent models within this band as tied. **Coverage.** API refusals (GPT models on viral or pathogen sequences) and empty responses (Sonnet at large N) reduce some cells; we report per-cell coverage with every number and omit the single most affected cell (Sonnet at N = 100). **Contamination.** As §4.8 establishes, dataset recognition is common but produces no detectable score-corrupting leakage (recognized assays score no higher than unrecognized ones, within ∼0.1 ρ), so memorized exposure does not measurably inflate the scores. The benchmark's resolution and coverage are therefore characterized, and contamination is bounded rather than assumed away.

5 Discussion

Substantive versus cosmetic reasoning. A small set of moves is substantive: retrieval of a named, checkable biochemical fact, and explicit, correct reasoning about assay polarity. These produce the field's best LLM traces and share a property, pinning a specific variant to a specific mechanism the assay rewards. Everything else (physicochemical severity, generic structure-breaking, mutation counting) generates the surface form of expertise without committing to anything that discriminates within the candidate set. The mechanistic probe quantifies the gap: roughly half of an LLM's ranking variance is a six-feature substitution-severity heuristic, so the elaborate structural narratives are, in aggregate, that heuristic with a rationale attached.

The rationale is not evidence. The most consequential finding for anyone tempted to deploy these models is that an LLM rationale carries little information about whether its ranking is correct, a domain-grounded instance of chain-of-thought unfaithfulness[21,22]. Confident, checkable-sounding specificity appears identically in correct traces and fabricated ones, so a reader cannot use the presence of a named motif as a quality signal. The structural story decorates a prediction that is mostly driven elsewhere, which is why the same machinery is anti-correlated on one protein and mildly predictive on the next.

What would have to change. Three things. First, grounding: replace confabulated structure with real retrieval, so that specificity tracks fact rather than fluency. Second, assay-conditioning: force explicit, validated reasoning about selection polarity before ranking, since the high-fitness and gain-of-function failures are direction errors, not knowledge errors. Third, a genuine interaction model to replace the additive backbone, which is precisely the part the specialized baselines already do and the LLM does not. Until then, an LLM's structural narrative is best treated as a hypothesis to verify, not as evidence that its ranking is correct.

6 Limitations

Our scope is substitution DMS in ProteinGym v1.3; indels and whole-proteome design are out of scope. The reasoning analyses cover only models that expose reasoning text (Gemini, Claude Opus, GLM); OpenAI hides raw chain-of-thought, so GPT participates via provider summaries, and Sonnet returned mostly empty visible text. The recognition rates are therefore confounded by reasoning-channel visibility, which we state rather than correct, and the audit detects stated recall, not silent use. The trace audit also relies on a single LLM judge (Gemini 3.5 Flash, itself a top-tier model) and is not validated against a second-family judge or human coding, so the strategy taxonomy (Figure 4a) and recognition rates should be read as exploratory. Each trace also receives a single dominant-strategy label, so traces that blend strategies are forced into one bucket; and because providers favour different strategies, the pooled per-strategy ρ partly reflects model mix, which Figure 4a shows split by provider. Two evaluation choices also bound our claims. We elicit a single joint ranking; we did not test pointwise or pairwise per-variant elicitation, so the set-size collapse (§4.2) is a property of joint-ranking elicitation and may overstate a pointwise model's limit. And we score with Spearman ρ over the (stratified) pool; a selection-relevant metric such as recall of the true top decile when picking K = 10 to 20 may rank methods differently, though the large ρ gap to specialists makes a reversal unlikely. Finally, ProteinGym fitness is itself measured with assay-specific noise, which upper-bounds any predictor's attainable ρ.

7 Conclusion

PG-Agent turns a one-off observation into a controlled, baseline-matched measurement. Frontier LLMs rank protein variants from sequence alone above chance, but they sit in the bottom quartile of specialized predictors, barely complement them, and lean on an over-applied substitution-severity heuristic; the regimes that matter most for design, large candidate pools and high-fitness shortlists, are exactly where every method, not just LLMs, loses signal. Their reasoning recognizes the benchmark often without corrupting its scores, and it does not signal its own correctness. That makes today's frontier LLMs strong collaborators for hypothesis articulation and weak substitutes for a fitness oracle.

References

1. Notin, P. et al. ProteinGym: large-scale benchmarks for protein fitness prediction and design. *NeurIPS Datasets & Benchmarks* (2023).
2. Frazer, J. et al. Disease variant prediction with deep generative models of evolutionary data (EVE). *Nature* 599, 91–95 (2021).
3. Laine, E., Karami, Y. & Carbone, A. GEMME: a simple and fast global epistatic model. *Mol. Biol. Evol.* 36, 2604–2619 (2019).
4. Meier, J. et al. Language models enable zero-shot prediction of the effects of mutations (ESM-1v). *NeurIPS* (2021).
5. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure (ESM-2). *Science* 379, 1123–1130 (2023).
6. Hsu, C. et al. Learning inverse folding from millions of predicted structures (ESM-IF1). *ICML* (2022).
7. Su, J. et al. SaProt: protein language modeling with structure-aware vocabulary. *ICLR* (2024).
8. Notin, P. et al. TranceptEVE: combining family-specific and family-agnostic models. *NeurIPS* (2022).
9. Tan, Y. et al. VenusREM: retrieval-augmented protein variant-effect prediction. (2024).
10. Anthropic. Claude Opus 4.8 system card: the ProteinGym-Hard (PG-Hard) variant-ranking evaluation. (2026).
11. Fowler, D. M. & Fields, S. Deep mutational scanning. *Nat. Methods* 11, 801–807 (2014).
12. Hopf, T. A. et al. Mutation effects predicted from sequence co-variation (EVmutation). *Nat. Biotechnol.* 35, 128–135 (2017).
13. Riesselman, A. J. et al. Deep generative models of genetic variation (DeepSequence). *Nat. Methods* 15, 816–822 (2018).
14. Rao, R. et al. MSA Transformer. *ICML* (2021).
15. Notin, P. et al. Tranception: protein fitness prediction with autoregressive transformers and retrieval. *ICML* (2022).
16. Madani, A. et al. Large language models generate functional protein sequences (ProGen). *Nat. Biotechnol.* 41, 1099–1106 (2023).
17. Yang, K. K. et al. Convolutions are competitive with transformers for protein modeling (CARP). (2022).
18. Li, M. et al. ProSST: protein language modeling with quantized structure. *NeurIPS* (2024).
19. Sun, W. et al. Is ChatGPT good at search? LLMs as re-ranking agents (RankGPT). *EMNLP* (2023).
20. Liu, N. F. et al. Lost in the middle: how language models use long contexts. *TACL* (2024).
21. Turpin, M. et al. Language models don't always say what they think. *NeurIPS* (2023).
22. Lanham, T. et al. Measuring faithfulness in chain-of-thought reasoning. arXiv:2307.13702 (2023).
23. Spearman, C. The proof and measurement of association between two things. *Am. J. Psychol.* 15, 72–101 (1904).

Supplementary

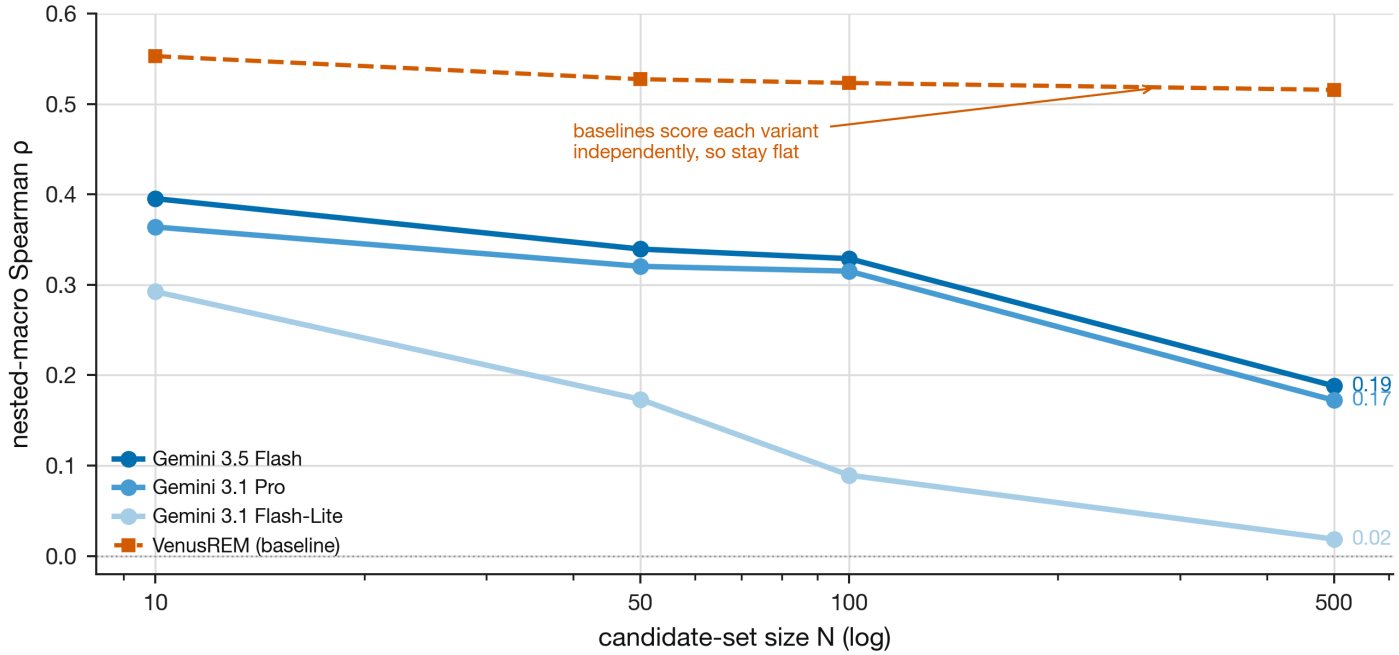

Figure S1 | The set-size collapse. Nested-macro ρ versus candidate-set size N (log axis) for three Gemini models out to N = 500, against the flat VenusREM reference. Baselines score each variant independently and stay flat, while every LLM degrades; the weakest collapses fastest, reaching ρ = 0.02 at N = 500.

Table S1 | Nested-macro Spearman ρ by model and set size (± SEM across the three runs). Bold = best LLM per column; single/multi = single- and multi-mutant ρ at N=50. Reduced coverage comes from API refusals (GPT, viral sequences) or empty responses (Sonnet at large N); Sonnet 4.6 at N=100 scored too few assays to report.

Model	ρ N=10 ± SEM	ρ N=50 ± SEM	ρ N=100 ± SEM	single ρ	multi ρ
Gemini 3.5 Flash	**0.395 ± 0.010**	**0.339 ± 0.007**	**0.329 ± 0.007**	0.342	0.310
Gemini 3.1 Pro	0.364 ± 0.022	0.320 ± 0.002	0.315 ± 0.002	0.324	0.297
GPT-5.5	0.353 ± 0.001	0.317 ± 0.011	0.317 ± 0.004	0.314	0.318
Claude Opus 4.8	0.382 ± 0.015	0.308 ± 0.009	0.301 ± 0.006	0.290	0.331
Claude Sonnet 4.6	0.341 ± 0.011	0.308 ± 0.023	n/a	0.287	0.307
Claude Opus 4.7	0.365 ± 0.016	0.284 ± 0.013	0.284 ± 0.006	0.253	0.311
GPT-5.4 mini	0.282 ± 0.023	0.225 ± 0.018	0.204 ± 0.004	0.218	0.208
GLM-5.2	0.254 ± 0.015	0.187 ± 0.008	0.125 ± 0.011	0.165	0.203
Gemini 3.1 Flash-Lite	0.292 ± 0.008	0.173 ± 0.010	0.089 ± 0.014	0.162	0.155
GPT-5.4 nano	0.230 ± 0.014	0.150 ± 0.028	0.114 ± 0.023	0.131	0.131
VenusREM (best of 95 baselines)	*0.553*	*0.527*	*0.523*	*0.511*	*0.516*

Table S2 | Nested-macro Spearman ρ by functional category (N = 50). Bold = best LLM per category. Stability and Expression are most tractable; Binding and Organismal Fitness hardest. VenusREM (best baseline) for reference.

Model	Activity	Binding	Expression	Org. Fitness	Stability
Gemini 3.5 Flash	0.34	0.30	0.37	**0.28**	**0.41**
Gemini 3.1 Pro	0.35	0.26	0.35	0.27	0.37
GPT-5.5	**0.36**	0.24	**0.38**	0.25	0.36
Claude Opus 4.8	0.34	0.27	0.34	0.22	0.37
Claude Sonnet 4.6	0.30	**0.37**	0.29	0.22	0.37
Claude Opus 4.7	0.31	0.26	0.32	0.20	0.33
GPT-5.4 mini	0.24	0.17	0.26	0.15	0.30
GLM-5.2	0.20	0.14	0.22	0.13	0.24
Gemini 3.1 Flash-Lite	0.19	0.13	0.24	0.12	0.19
GPT-5.4 nano	0.15	0.09	0.21	0.08	0.21
VenusREM (best baseline)	*0.49*	*0.45*	*0.57*	*0.46*	*0.65*

Table S3 | Nested-macro Spearman ρ by functional category at N = 10. Bold = best LLM per category.

Model	Activity	Binding	Expression	Org. Fitness	Stability
Gemini 3.5 Flash	**0.45**	0.35	**0.44**	0.27	**0.47**
Gemini 3.1 Pro	0.36	0.33	0.42	**0.28**	0.42
GPT-5.5	0.39	0.29	0.42	0.22	0.44
Claude Opus 4.8	0.40	0.34	0.43	0.27	0.47
Claude Sonnet 4.6	0.38	0.29	0.38	0.24	0.42
Claude Opus 4.7	0.38	**0.41**	0.38	0.22	0.44
GPT-5.4 mini	0.34	0.22	0.34	0.17	0.33
GLM-5.2	0.30	0.20	0.25	0.19	0.33
Gemini 3.1 Flash-Lite	0.32	0.25	0.32	0.22	0.36
GPT-5.4 nano	0.25	0.18	0.28	0.18	0.27

Table S4 | Nested-macro Spearman ρ by functional category at N = 100. Bold = best LLM per category. Claude Sonnet 4.6 omitted (too few assays scored at N=100).

Model	Activity	Binding	Expression	Org. Fitness	Stability
Gemini 3.5 Flash	0.35	**0.30**	0.34	**0.27**	**0.38**
Gemini 3.1 Pro	0.33	0.27	0.35	0.26	0.36
GPT-5.5	**0.36**	0.27	**0.38**	0.26	0.33
Claude Opus 4.8	0.33	0.26	0.32	0.22	0.37
Claude Opus 4.7	0.30	0.29	0.30	0.19	0.33
GPT-5.4 mini	0.21	0.15	0.22	0.16	0.27
GLM-5.2	0.12	0.12	0.12	0.10	0.17
Gemini 3.1 Flash-Lite	0.09	0.09	0.11	0.05	0.09
GPT-5.4 nano	0.11	0.07	0.13	0.06	0.19

Built on the ProteinGym corpus (Marks Lab / OATML); the variant-ranking framing follows the PG-Hard evaluation in the Claude Opus 4.8 system card.