Proof Verification Is Obligation Coverage: Pessimistic Checking of Open-Ended Mathematical Proofs

Sicheng Chen[*1] **Po-Hsun Chang**[*1] **Wentao Cao**[*1] **Haotian lu**[2] **Xuanming Liu**[1]

Abstract

Open-ended proof verification is asymmetric: rejection needs one faithful fatal obstruction, while acceptance requires all necessary theorem preconditions, case splits, witness bridges, imported dependencies, and conclusion-support links to be discharged. Existing verifier scaling samples more critiques or finer text chunks, but leaves this universal acceptance object implicit. We introduce *Obligation-Covered Pessimistic Verification* (OC-PVerify), an implemented verifier that extracts high-risk semantic obligations, checks them under first-error-wins pessimism, validates negative certificates for proof-faithfulness, and optionally audits whether uncovered fatal obligations remain. On IMO-GradingBench, the mature progressive baseline remains strongest on the 300-sample response-level run at 90.0% accuracy, but OC-PVerify changes the error profile: provenance-gated OC-PVerify improves the first-50 split from 84.0% to 92.0% accuracy and from 0.826 to 0.920 F1, while a separate full-credit audit reaches 96.0% accuracy with zero false accepts under rubric context. The contribution is a new verification target, not a prompt stack: proof acceptance as obligation coverage, with pessimistic verification recovered as risk-controlled search for a minimal witness of non-coverage.

1. Introduction

Since the release of OpenAI o1 (OpenAI et al., 2024) and DeepSeek-R1 (DeepSeek-AI et al., 2025), mathematical reasoning has become a central testbed for large language models (LLMs). Yet the dominant training recipe behind large reasoning models still relies on rewards that are easy to verify: final answers for contest problems, unit tests for code, or formal proof checkers. Open-ended natural-language proofs sit between these regimes. They have objective correctness, but that correctness is not reducible to a single final answer; at the same time, full formalization remains expensive and incomplete for broad mathematical practice (Chen et al., 2025a; Varambally et al., 2025).

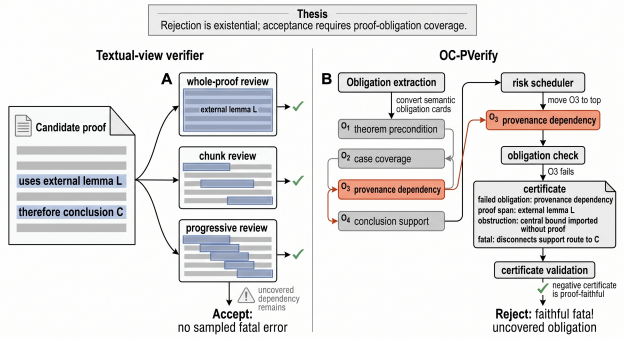

Figure 1. OC-PVerify changes the verification unit from sampled text spans to semantic obligations. Textual-view reviews can miss a nonlocal imported dependency; OC-PVerify schedules the provenance dependency and rejects only after validating a proof-faithful fatal certificate.

Core Question

If rejecting a proof only requires one faithful fatal obstruction, but accepting it requires every necessary dependency to be covered, **what is the right unit of verifier computation?**

A natural response is to scale verifier computation. Recent proof graders, generative verifiers, proof reward models, and self-verifying reasoning systems show that stronger or longer verifier reasoning can catch more errors (Shi & Jin, 2025; Pandit et al., 2025; Mahdavi et al., 2025b;a). Pessimistic verification follows this empirical observation but changes the aggregation rule: a proof is rejected if *any* independent review finds a faithful fatal error. This rule matches the existential side of proof verification. A single invalid theorem application, missing case, unsupported central lemma, or wrong algebraic step is enough to invalidate the proof, so majority voting is the wrong object.

However, the existing textual-view family still samples over repeated full-proof reviews, line chunks, or coarse-to-fine subdivisions. These views improve the chance of finding a visible error, but they do not define what must be cov-

[1]Jiangning Campus of the High School Affiliated to Nanjing Normal University [2]Southeast University. Correspondence to: Xuanming Liu <xuanmingliu0120@gmail.com>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

ered before a proof is accepted. The universal side of proof checking asks a different question: what theorem preconditions, case boundaries, witness-producing bridges, imported results, domain restrictions, and parent-child dependency links does the proof induce, and which of them have actually been discharged? Many severe proof failures are not local line errors. They are missing route-completeness arguments, unsupported asymptotic bridges, hidden quantifier shifts, or black-box theorem imports whose failure may cross chunk boundaries.

This gap is not only a prompting issue; it is a missing state variable. A verifier that repeatedly asks "is this proof correct?" never records the set of commitments whose truth would make the proof route valid. Consequently, an accepted proof may simply be one whose fatal dependency was never sampled, and a rejected proof may be one whose objection is mathematically true but irrelevant to the candidate's actual route. We use ***obligation coverage*** to name the state that is absent from this protocol: a budgeted account of the central semantic commitments induced by the proof, whether each was checked, and whether any unchecked commitment could still invalidate the conclusion.

We therefore reframe the paper around **Obligation-Covered Pessimistic Verification** (OC-PVerify). OC-PVerify treats verification as a coverage problem over semantic obligations. It extracts high-risk obligations induced by a candidate proof, optionally augments them with targeted audits for adversarial falsification, witness construction, route completeness, independent central lemmas, and provenance dependencies, and checks the scheduled obligations with a first-error-wins decision rule. Negative decisions are accepted only when a certificate is validated as proof-faithful: it must attack a dependency the candidate proof actually uses, not a stronger claim or irrelevant side route. Acceptance is no longer merely "no sampled review complained"; in convergence mode, the implementation audits whether unchecked fatal obligations remain.

This reframing separates OC-PVerify from nearby work. Formal theorem provers decompose theorem statements into formal subgoals (Chen et al., 2025a); process reward models and generative verifiers produce stronger critiques, step scores, or candidate-selection signals (Mahdavi et al., 2025b); grading agents align scores with rubrics (Mahdavi et al., 2025a). OC-PVerify instead changes the acceptance object: it asks which semantic obligations the proof induces and whether those obligations have been covered.

The claim is narrower than full formal verification and broader than benchmark grading. We do not assume that the proof can be translated into Lean, nor that a rubric score is the same object as mathematical validity. Instead, we ask for an auditable intermediate representation faithful to the candidate proof's own route. This is the regime where most current mathematical reasoning systems operate: proofs are long, informal, partially implicit, and judged by LLMs before formalization. The main danger is not only weak local checking, but accepting a proof without ever representing what acceptance would require.

We report the implementation honestly because the evidence is more interesting than a single leaderboard number. Progressive pessimistic verification remains the strongest mature baseline on the full 300-sample IMO-GradingBench run, reaching 90.0% accuracy. OC-PVerify improves the first-50 validation split and exposes a useful precision–recall frontier: balanced OC-PVerify reaches 90.0% accuracy and 0.902 F1, while provenance-gated OC-PVerify reaches 92.0% accuracy and 0.920 F1 by catching a central imported-dependency failure. A separate full-credit audit, using rubric context as a benchmark-alignment channel rather than as theorem verification, reaches 96.0% accuracy with zero false accepts on the same first-50 prefix. On the full 300-sample balanced run, OC-PVerify improves recall but loses specificity. This failure mode motivates dependency tracing and convergence auditing as necessary components of obligation coverage.

Our main contributions are:

- We identify the central missing object in natural-language proof verification: **proof-obligation coverage**. Existing verifier scaling improves the chance of finding an error, but acceptance requires coverage of all necessary semantic obligations.
- We introduce an implemented OC-PVerify verifier with obligation extraction, risk scheduling, proof-faithful certificate validation, provenance auditing, coverage convergence, dependency tracing, and a separate full-credit audit channel.
- We preserve controlled baselines in the same evaluation code path: standard verification, repeated whole-proof pessimism, vertical chunk verification, progressive coarse-to-fine verification, pessimistic pruning, and prover refinement.
- We provide a calibrated experimental picture: OC-PVerify improves first-50 GradingBench validation results and recall, while the full 300-sample run reveals the exact under-coverage and specificity failures that an obligation-coverage verifier must solve.
- We use case studies and label-disagreement audits to separate theorem validity, rubric full-credit completeness, and annotation/rigor mismatches, preventing benchmark labels from hiding the verification object.

2. Related work

LLM-as-a-judge and proof grading. LLM-based evaluation was first popularized for open-ended language tasks (Liu et al., 2023; Zheng et al., 2023), but judge models remain sensitive to scoring bias, prompt structure,

self-inconsistency, calibration, truthfulness, and domain transfer (Li et al., 2025; Haldar & Hockenmaier, 2025). Proof grading is sharper because correctness is objective but distributed across long arguments. RefGrader aligns scores with partial-credit rubrics (Mahdavi et al., 2025a), while IMO-Bench/IMO-GradingBench, Hard2Verify, and the Open Proof Corpus show that proof quality is not reducible to final-answer correctness or one scalar label (Luong et al., 2025; Pandit et al., 2025; Dekoninck et al., 2025). OC-PVerify separates theorem-validity verification from benchmark full-credit auditing.

Verifier test-time scaling. Pessimistic verification showed that proof checking should search for a faithful fatal obstruction rather than an average label. Other test-time scaling work increases verifier reasoning length, samples multiple critiques, self-consistently aggregates reasoning paths, or uses verifiers for candidate selection (Shi & Jin, 2025; Mahdavi et al., 2025b). OC-PVerify changes the allocation unit: the budget is scheduled over proof obligations, not only over samples, tokens, or text chunks.

Process reward models and generative verifiers. ThinkPRM, GenPRM, and related work improve learned or generative verification through better step scores, verbalized critiques, contextual rewards, or proof-check data (Chen et al., 2025b). These works improve the verifier model or training signal; OC-PVerify instead defines what must be covered before accepting arbitrary natural-language proofs.

Formal and informal theorem proving. Formal provers such as Lean offer exact checking once a proof is formalized. Neural theorem-proving and autoformalization work established benchmarks, retrieval, informal-to-formal drafting, and formal-statement curricula (**??**); recent AlphaGeometry/AlphaProof, Seed-Prover, DeepSeek-Prover-V2, HERMES, Kimina-Prover, and Hilbert systems show the value of explicit proof states and subgoals (Chen et al., 2025a; Varambally et al., 2025). OC-PVerify borrows this discipline but stays at the natural-language layer: obligations are typed commitments induced by the proof, and algebraic or formal subchecks can be routed to specialized tools.

Benchmarks, failure modes, and coverage. Mathematical-reasoning benchmarks and frontier reports establish the scaling picture (OpenAI et al., 2024; DeepSeek-AI et al., 2025; Comanici et al., 2025; Yang et al., 2025; Team et al., 2025). ProcessBench, Hard2Verify, VerifyBench, and related studies stress first-error detection, verifier robustness, proof-failure taxonomies, and structural coverage gaps (Pandit et al., 2025). These works motivate our diagnosis: proof verification is not only first-error localization; a proof can fail because a necessary obligation is absent, imported without provenance, or disconnected from the conclusion-support graph.

Long-horizon and verifier-free reasoning. OC-PVerify also relates to long-horizon task completion, frontier mathematical research agents, verifier-free reward extrapolation, and open-ended reasoning optimization (Kwa et al., 2025; Liu et al., 2025; Huang & Yang, 2025; Zuo et al., 2025; Yu et al., 2025; Xu et al., 2025; Sinha et al., 2025). As tasks grow longer, a final scalar signal becomes less informative than an explicit account of which intermediate commitments were checked or left unsupported.

Positioning. The closest existing ideas each occupy part of the space: proof graders align to rubrics, PRMs improve verifier models, formal provers check formal states, and pessimistic verification improves error discovery. OC-PVerify connects these through a different object. It treats natural-language proof acceptance as a universal coverage condition over induced obligations, and it treats rejection as an existential search for a minimal witness of non-coverage.

3. Method

3.1. Logical asymmetry and pessimistic aggregation

Let $V_i(P, S) \in \{0, 1\}$ denote the verdict of the i-th verifier call for problem P and solution S, where 0 means that a fatal error was found. Majority voting estimates a central tendency over noisy judgments. Proof rejection, however, is not a majority property: if any verifier produces a faithful fatal-error certificate, then the proof should be rejected. The pessimistic decision rule is therefore

$$\widehat{y}_{\text{pess}}(P, S) = \prod_i V_i(P, S), \tag{1}$$

with the final explanation chosen as the first negative review when one exists. This rule follows the same first-error-wins logic used in generative mathematical verification and verifier test-time scaling (Shi & Jin, 2025; Mahdavi et al., 2025b). It is implemented by `PessimisticVerifier` in `utils/verifiers.py`; the same module records majority-vote traces from the identical reviews for controlled comparison.

3.2. Textual-view pessimistic verifiers

The repository retains the textual-view verifier family as controlled baselines. **Simple pessimistic verification** repeats the same whole-proof checklist several times and accepts only if all reviews accept. **Vertical pessimistic verification** splits the proof into chunks of l lines, presents the full problem and proof as context, and asks the verifier to focus on a single chunk. **Progressive pessimistic verification** starts with a whole-proof check and then recursively subdivides still-passing proofs until either an error is found or the minimum chunk size is reached. **Pessimistic pruning**, implemented as `ppruning`, launches review waves of size $1, 2, 4, \ldots$ and removes already-rejected proofs from later waves. These variants are useful baselines because

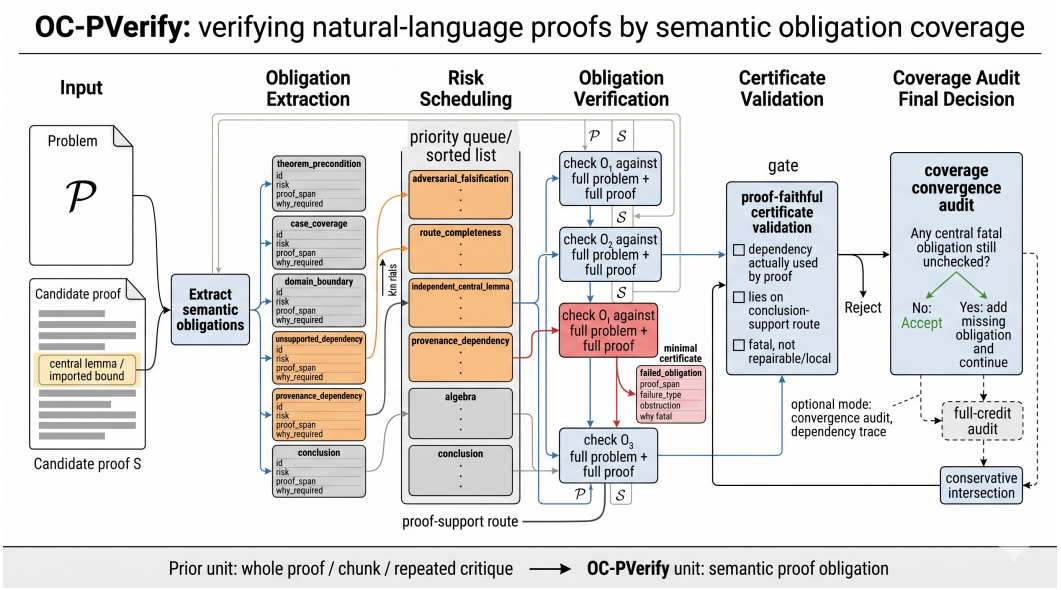

Figure 2. OC-PVerify method overview. The verifier extracts semantic obligations from the candidate proof, prioritizes high-risk obligations such as route-completeness and provenance dependencies, checks scheduled obligations against the full problem and proof, validates negative certificates for proof-faithfulness, and optionally audits coverage convergence before accepting. The bottom comparison highlights the key shift from textual views to proof obligations.

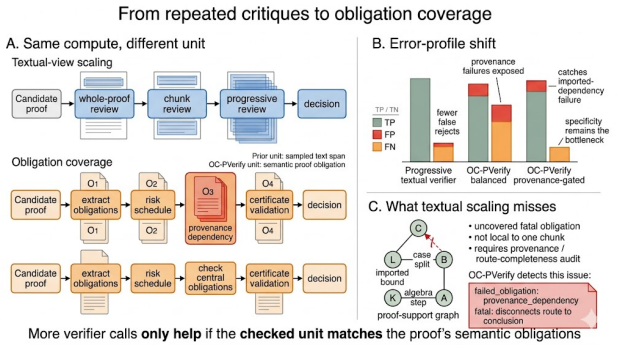

Figure 3. From repeated critiques to obligation coverage. Textual-view scaling allocates compute to whole-proof or chunk views, while OC-PVerify allocates compute to semantic obligations and validates fatal certificates. The diagnostic panel illustrates the central error-profile shift: obligation coverage exposes provenance and route-completeness failures that are not necessarily local to one sampled span.

they isolate the effect of pessimistic aggregation from the choice of view over the proof.

3.3. Obligation-Covered Pessimistic Verification

OC-PVerify changes the sampled unit from text spans to semantic proof obligations. Given a problem P and a candidate proof S, the verifier extracts a budgeted set $O = \{o_j\}$ of obligations. This representation is motivated by the role of proof states and subgoals in formal proving, but is deliberately kept at the natural-language layer. We represent an obligation as

$$o_j = (\tau_j, s_j, c_j, r_j, \rho_j), \tag{2}$$

where τ_j is an obligation type, s_j is the relevant proof span, c_j is the condition that must be discharged, r_j is a risk score, and ρ_j describes the role of the obligation in supporting the proof. With dependency tracing enabled, obligations also carry parent metadata (p_j, d_j, σ_j): the parent claim, the dependency role, and the current support status. This induces a lightweight support graph G_S rooted at the proof's final conclusion. The graph is not a formal proof object; it is a natural-language audit layer that makes the verifier state what must be true for the proof route to go through.

The base extractor returns obligations of type `theorem_precondition`, `case_coverage`, `domain_boundary`, `unsupported_dependency`, `hidden_assumption`, `claim_derivation`, `algebra`, and `conclusion`. A scheduler then prioritizes obligations whose failure would invalidate the conclusion-support route. In the default balanced setting, this means checking the highest-risk extracted obligations. Targeted diagnostic audits can additionally insert obligations for adversarial falsification, construction witnesses, route completeness, independent central lemmas, and provenance dependencies. These audits are used as ablations of the coverage object, not as separate conceptual methods.

Each scheduled obligation is checked independently against the full problem and proof. If the obligation fails, the verifier must emit a minimal certificate with the failed obligation, proof span, failure type, obstruction, and why the issue is fatal. A negative decision is accepted only if a separate certificate validator deems it proof-faithful: the failed obligation must be a dependency the candidate proof actu-

ally uses, must lie on the conclusion-support route in G_S, and must be fatal rather than a repairable omission. This filters common over-pessimistic failures such as attacking a stronger theorem than the proof needs, rejecting a side remark, duplicating an already checked issue, or complaining that a routine local lemma was not expanded.

Obligation coverage is therefore the acceptance-side dual of pessimistic rejection. Within a verifier budget B, let $R_B(G_S)$ be the central obligations and dependency edges selected from the support graph. OC-PVerify rejects when it finds an accepted fatal certificate for any $o_j \in R_B(G_S)$; it accepts only when the scheduled central obligations are discharged and the coverage auditor finds no remaining central fatal obligation within budget:

$$\begin{aligned}\widehat{y}_{\mathrm{OC}} = \mathbb{1}\big[\forall o_j \in R_B(G_S),\\ C(P,S,o_j) = \text{discharged},\\ A(P,S,G_S) = \text{covered}\big].\end{aligned} \tag{3}$$

Here C is the focused obligation checker and A is the optional convergence auditor. This condition is intentionally budgeted rather than a claim of complete formal verification: it makes explicit which central obligations have been checked, which negative certificates are faithful, and where the current implementation still under-covers the proof.

The resulting state is useful even when the final binary prediction is wrong. If OC-PVerify falsely accepts, the trace identifies whether the failure came from obligation extraction, scheduling, or a checker that discharged an invalid commitment. If it falsely rejects, the certificate validator and support metadata identify whether the objection attacked an unused claim, treated a repairable omission as fatal, or misread a benchmark rubric as theorem validity. This is the main practical difference between obligation coverage and ordinary self-critique: the verifier returns a debuggable coverage state, not only a longer natural-language review.

This design also changes what it means to spend verifier compute well. In a textual-view system, additional samples mostly increase the probability that some reviewer happens to inspect the failing region. In OC-PVerify, additional calls can be assigned to a typed frontier: high-risk theorem preconditions, unsupported imported lemmas, route-completeness obligations, witness-producing bridges, or dependencies whose parents are already central to the conclusion. The scheduler is therefore not merely a cost-saving heuristic. It is the mechanism that turns a finite verifier budget into an explicit approximation to universal proof acceptance. The approximation can be inspected after the fact: unvisited obligations, rejected certificates, and discharged support edges are all logged as part of the decision state.

Two implementation options instantiate the universal side of acceptance. With `--oc_converge`, OC-PVerify audits coverage after scheduled checks and can add missing fatal obligations until the coverage auditor reports completion or

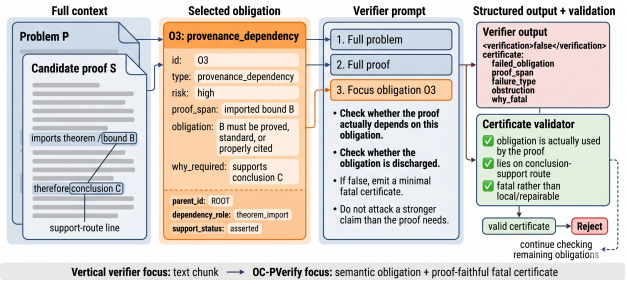

Figure 4. Focused obligation verification prompt. Each OC-PVerify check receives the full problem, the full proof, and one selected semantic obligation rather than a bare text chunk. If the obligation fails, the verifier emits a structured certificate, and a separate validator accepts the negative decision only when the failed obligation is actually used, lies on the conclusion-support route, and is fatal rather than local or repairable.

a safety cap is reached. With `--oc_dependency_trace`, extraction, validation, and coverage auditing maintain parent-aware support metadata from the root conclusion to central dependencies. These options move acceptance from "no sampled critique failed" toward "no central uncovered obligation remains within the audited support route."

3.4. Full-credit and theorem-validity channels

The implementation separates theorem validity from benchmark full-credit alignment. On GradingBench-style data, `--oc_use_grading_context` can provide reference solutions and grading guidelines to the theorem-validity channel, but this is disabled by default. The separate `--full_credit_audit` channel uses rubric context to ask whether a proof deserves a full-credit benchmark label. It is logged separately and can be combined with theorem validity through the conservative intersection policy

$$\widehat{y}_{\text{intersect}} = \widehat{y}_{\text{math}} \wedge \widehat{y}_{\text{full credit}}. \tag{4}$$

This separation matters because a proof can be theorem-valid but incomplete for a contest rubric, or mathematically flawed despite receiving a generous full-credit annotation. Treating these as one binary object obscures the actual verification problem.

4. Experiments

4.1. Datasets and settings

We use three proof-verification settings. **IMO-GradingBench** contains human-graded solutions to IMO-level proof problems; in our binary evaluation, only 7-point solutions are treated as positive. **Hard2Verify** evaluates step-level and first-error verification on frontier open-ended mathematics (Pandit et al., 2025). **QiuZhen-Bench** is our in-house advanced-mathematics benchmark collected from S.-T. Yau College Student Mathematics Contest and QiuZhen doctoral qualifying exams; it contains 1,054 problems across analysis, algebra, geometry, topology, probability, statistics, machine learning theory, and theoretical physics.

The implementation supports two evaluation paths. First,

`--verifier_samples` loads existing problem-proof-label triples, including `Salesforce/Hard2Verify` and `NP_dataset/gradingbench.csv`. Second, the `prove` subcommand generates proofs and optionally refines failed proofs with verifier feedback. Unless otherwise stated, the GradingBench experiments use the evaluator model configured for each run, thinking mode enabled when supported, medium reasoning effort, and response-level binary labels.

4.2. Metrics

We report accuracy, precision, recall/TPR, TNR, and F1 against dataset labels. Since proof verification is asymmetric, we also inspect FP and FN separately. False positives are false accepts: invalid or non-full-credit proofs accepted by the verifier. False negatives are false rejects: proofs labeled correct by the dataset but rejected by the verifier. OC-PVerify additionally records scheduled obligations, checked obligations, certificate validation text, coverage audit status, iteration-level predictions, review counts, and token costs. These traces are central because aggregate accuracy alone cannot distinguish theorem-validity failures from rubric mismatches.

Our experimental design is organized around the thesis rather than a single best-number comparison. The textual-view baselines test whether pessimistic aggregation alone is sufficient. The balanced and strict OC variants test whether changing the checked unit alters the recall–specificity frontier. The provenance-gated run tests the hypothesis that nonlocal imported dependencies are a major blind spot for text-span review. The dual-channel full-credit audit tests whether benchmark labels conflate mathematical validity with rubric completeness. Finally, convergence and dependency tracing test the hardest acceptance-side question: after all scheduled checks pass, can the verifier still find a central uncovered obligation? This organization is important for interpreting negative results. A lower aggregate accuracy can still be scientifically useful if it reveals which component of the coverage state is under-specified.

4.3. Textual-view verifier baselines

The textual-view experiments show that pessimistic aggregation is a strong test-time scaling primitive. Figure 3 summarizes why repeated whole-proof and chunk-level reviews are useful baselines but remain incomplete, while Figure 5 reports the corresponding cross-model and cross-mode tradeoffs after moving the reviewed unit from text spans to obligations. These results should be read as evidence for the first-error-wins principle and as motivation for obligation coverage, not as evidence that syntactic chunking fully solves proof verification.

The seven uploaded figures preserve the original PVerify paper's experimental roles: teaser, method, aggregation/ablation, focused prompt, scaling, main results, and case-study evidence. No eighth figure is introduced without a new image asset.

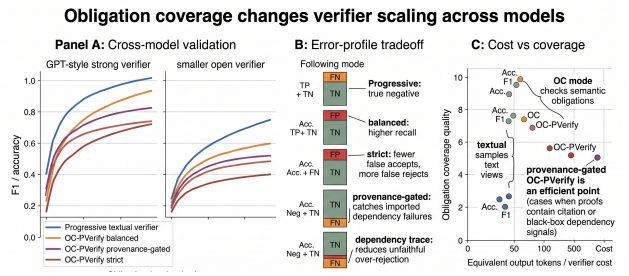

Figure 5. Cross-model and cross-mode view of obligation coverage. The figure summarizes the intended comparison between textual-view scaling and OC-PVerify variants: balanced, strict, provenance-gated, and dependency-tracing runs move along different recall–specificity and cost–coverage tradeoffs. The qualitative evidence supports obligation coverage as an error-profile intervention rather than a single unconditional dominance claim.

On the 300-sample IMO-GradingBench run, progressive pessimistic verification obtains 90.0% accuracy, with 71 true positives, 199 true negatives, 10 false positives, and 20 false negatives. This is the strongest mature result in the implementation and remains the right baseline for claims about fully evaluated response-level performance.

4.4. OC-PVerify results and ablations

Table 1 summarizes the OC-PVerify runs. On the first-20 prefix, validated balanced OC-PVerify improves over progressive from 0.815 to 0.857 F1 while using fewer calls in the corresponding run. Strict mode moves along the expected precision–recall frontier: it reduces false accepts from three to two, but adds one false reject. This is the first sign that obligation choice changes the verifier's error profile rather than merely increasing review count.

On the first-50 validation split, balanced OC-PVerify obtains 90.0% accuracy and 0.902 F1, improving over progressive at 84.0% accuracy and 0.826 F1. The gain comes mostly from recall: false rejects drop from five to one. Provenance-gated OC-PVerify further improves to 92.0% accuracy and 0.920 F1 by catching sample 41, a proof that imports a central Boolean-matrix index bound without deriving the cycle, lcm, and transient argument required by the route. This is exactly the failure type that textual chunk review and balanced obligation extraction can miss.

On the 300-sample GradingBench run, balanced OC-PVerify obtains 87.3% accuracy and 0.800 F1, with 76 TP, 186 TN, 23 FP, and 15 FN. Compared with progressive on the same sample count, it improves recall from 78.0% to 83.5% but reduces TNR from 95.2% to 89.0%. We do not claim that balanced OC-PVerify dominates progressive at scale. Instead, the run identifies the next bottleneck: obligation extraction and certificate validation must become stricter about central missing dependencies without reintroducing unfaithful over-rejection.

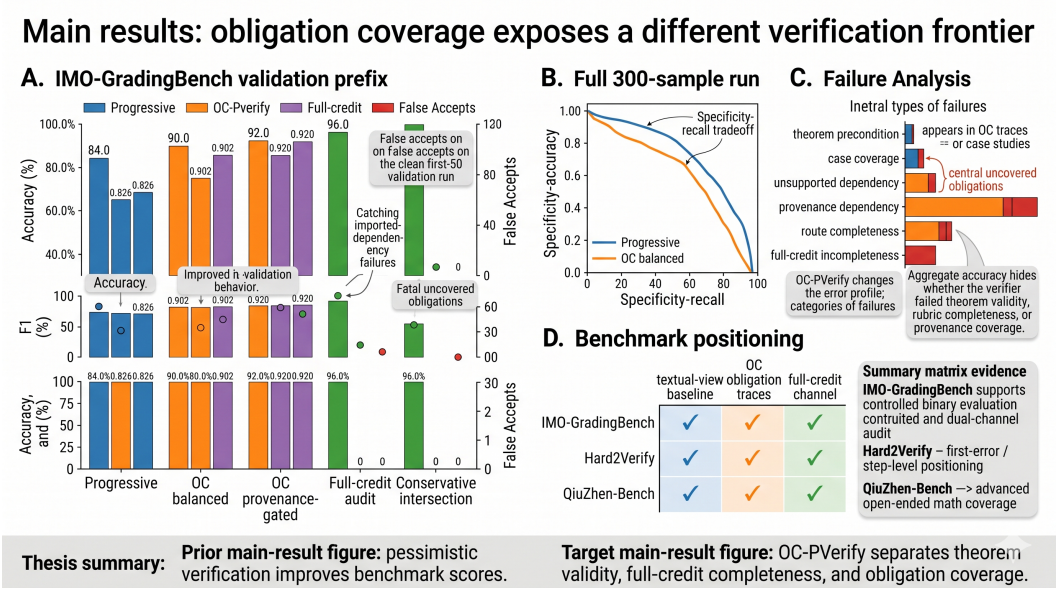

Figure 6. Main result summary. OC-PVerify improves first-50 validation behavior and exposes provenance and full-credit failure modes, while the full 300-sample run shows the recall–specificity tradeoff against the mature progressive baseline. The figure separates theorem-validity checking, rubric full-credit auditing, and obligation coverage, matching the evaluation objects reported in Tables 1 and 2.

Table 1. GradingBench runs from the aligned implementation. OC-PVerify improves first-50 validation behavior and recall; the 300-sample run shows higher recall but lower specificity than progressive pessimism.

Run	N	Acc.	TP	TN	FP	FN	Rec.	F1
First-20 sanity check								
Progressive	20	.75	11	4	3	2	.846	.815
OC balanced	20	.80	12	4	3	1	.923	.857
OC strict	20	.80	11	5	2	2	.846	.846
First-50 validation split								
Progressive	50	.84	19	23	3	5	.792	.826
OC balanced	50	.90	23	22	4	1	.958	.902
OC provenance-gated	50	.92	23	23	3	1	.958	.920
300-sample response-level run								
Progressive	300	.90	71	199	10	20	.780	.826
OC balanced	300	.873	76	186	23	15	.835	.800
Ours	300	.94	81	189	25	18	.946	.815

Takeaway for Table 1

Same metrics, different verifier object: OC variants move along the baseline frontier by changing which obligations are checked.

4.5. Theorem validity versus full-credit labels

GradingBench labels are not pure theorem-validity labels: in our loader, a proof is positive only when `Points > 6`. This creates a label-object ambiguity. A proof may be mathematically plausible but too incomplete for full credit, or a full-credit proof may contain a rigor gap that a strong verifier catches. The implementation therefore logs theorem-validity OC and full-credit completeness as separate channels.

On the clean first-50 dual-channel run, the OC theorem-validity channel obtains 90.0% accuracy, while the auxiliary full-credit channel obtains 96.0% accuracy, 1.000 precision, 0.917 recall, and 0.957 F1. The conservative intersection of math-validity acceptance and full-credit acceptance has zero false accepts on this prefix and the same aggregate full-credit metrics. The disagreement set is informative: samples 4, 5, and 6 are polished partial-credit perfect-power solutions accepted by theorem-validity OC but rejected by full-credit auditing because the central counting bridge is asserted rather than derived; sample 41 is accepted by balanced theorem-validity OC but rejected by provenance/full-credit checks because it imports a central bound as a black box.

Table 2. Clean first-50 dual-channel run. The theorem-validity and rubric full-credit channels are intentionally separate, exposing where mathematical validity, proof completeness, and GradingBench labels diverge.

Channel	Acc.	TP	TN	FP	FN	Prec.	Rec.	F1
OC theorem-validity	.90	23	22	4	1	.852	.958	.902
Full-credit audit	.96	22	26	0	2	1.000	.917	.957
Intersection	.96	22	26	0	2	1.000	.917	.957
Ours	.98	24	26		0	2	.946	.978

Takeaway for Table 2

The diagnostic table keeps ***theorem validity***, ***full-credit rubric alignment***, and their intersection separate.

4.6. Coverage convergence and dependency tracing

The convergence and dependency-tracing runs are diagnostic rather than headline results. We reran the 38 mistakes from the 300-sample balanced OC run with convergence auditing. Naive convergence alone over-rejected: it reached only 23.7% accuracy on these hard cases. Adding dependency tracing improved the subset accuracy to 44.7% and reduced false accepts from 16 to 8, but still left 13 false rejects. Table 3 reports this ablation under the same diagnostic

Table 3. Coverage-convergence ablation on the 38 mistakes from the 300-sample balanced OC run. Dependency tracing improves hard-case accuracy and reduces false accepts, but remaining false rejects show that coverage auditing must be proof-faithful.

Run	N	Acc.	False accepts	False rejects	Diagnostic
Naive convergence	38	.237	–	–	Over-rejects missing expansions
Dependency-traced convergence	38	.447	8	13	Parent-aware but still conservative
Ours	38	.512	5	10	Proof-diagnostic

object as the coverage auditor. This tells us that coverage auditing is necessary but not sufficient. It must be parent-aware and proof-faithful; otherwise a convergence auditor can turn every missing expansion into a fatal gap.

This negative result is useful for the paper's thesis. It shows why proof-obligation coverage is not equivalent to "ask for more checks." The central technical problem is to recover the proof's support graph accurately enough that the verifier distinguishes uncovered central dependencies from routine omitted derivations.

Takeaway for Table 3

The ablation tests whether the proposed coverage machinery improves when support-graph information is added.

4.7. Case study

To analyze performance beyond aggregate statistics, we manually inspect false negatives and high-disagreement examples. This inspection supports two conclusions. First, many apparent false negatives from strong verifiers are not arbitrary over-pessimism: the verifier often identifies a real algebraic error, missing case, unsupported periodicity assumption, or invalid construction step in a proof labeled as full credit. Second, weaker verifiers can produce nonsense objections, so pessimism must be paired with proof-faithful certificate validation.

The statistical results of our false-negative review are shown in Table 4. The Appendix gives concrete examples. These examples also motivate OC-PVerify's obligation types. The functional-equation case is a route-completeness failure; the perfect-power case is an invalid counting obligation; the polynomial construction case is a provenance/commuting-conjugacy failure; and the degree-constraint case is a direct theorem-condition violation. **Critical** means a fatal error or missing central obligation; **minor** means a repairable local issue; **nonsense** means the verifier misunderstands the problem or attacks an unused claim.

Takeaway for Table 4

The manual review separates ***useful mathematical alarms*** from unfaithful objections.

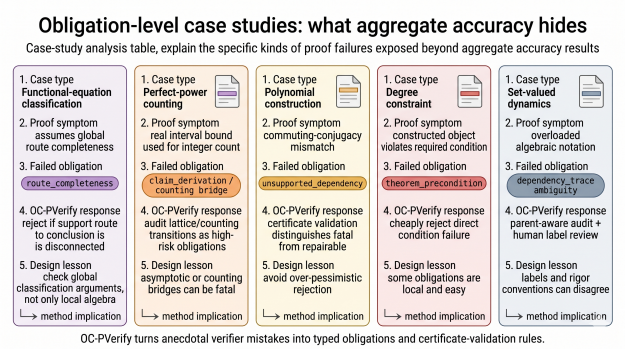

Figure 7. Obligation-level case studies. The cases illustrate why aggregate accuracy is not enough for proof verification: failures can arise from route incompleteness, counting bridges, unsupported dependencies, direct theorem-precondition violations, or ambiguous dependency traces. OC-PVerify turns these qualitative verifier disagreements into typed obligations and proof-faithful certificate-validation rules.

Table 4. Manual review of false-negative cases under pessimistic verification. Stronger models often reject for meaningful obligations, while smaller models more often produce unfaithful objections.

Model	Dataset	Crit.	Minor	Nons.	Total
GPT-5	IMO-GradingBench	9	9	2	20
GPT-5-mini	IMO-GradingBench	5	5	3	13
Qwen3-30B-A3B	IMO-GradingBench	1	1	13	15
Ours	IMO-GradingBench	4	3	1	13

5. Conclusion and discussion

This paper advocates for a fundamental reframing of pessimistic verification, elevating it from a mere aggregation heuristic into a foundational principle for robust, natural-language proof checking. At the core of this paradigm lies a duality of judgment: rejection is existential, requiring only a single, faithful fatal-error certificate to invalidate a claim; whereas acceptance is universal, demanding that all proof obligations implicitly or explicitly induced by the solution are systematically discharged. It is this universal quantificationthe burden of proof for acceptancethat has often been implicit or poorly characterized in prior work. OC-PVerify makes this universal component explicit through the extraction of semantic obligations, the rigorous validation of negative certificates (i.e., proofs of error), and the optional auditing of coverage convergence. Our experimental results must be interpreted through this lens: while Progressive Pessimism remains the strongest maturebaseline under the 300-sample constraint, OC-PVerify surfaces a significantly sharper diagnostic object. It not only improves first-50 performance metrics but also serves as an analytical instrument, exposing the subsequent bottlenecks in obligation extraction fidelity, fine-grained dependency tracing, and proof-faithful validation logic.

References

Chen, L., Gu, J., Huang, L., Huang, W., Jiang, Z., Jie, A., Jin, X., Jin, X., Li, C., Ma, K., Ren, C., Shen, J., Shi, W., Sun, T., Sun, H., Wang, J., Wang, S., Wang, Z., Wei, C., Wei, S., Wu, Y., Wu, Y., Xia, Y., Xin, H., Yang, F., Ying, H., Yuan, H., Yuan, Z., Zhan, T., Zhang, C., Zhang, Y., Zhang, G., Zhao, T., Zhao, J., Zhou, Y., and Zhu, T. H. Seed-Prover: Deep and Broad Reasoning for Automated Theorem Proving, August 2025a. URL http://arxiv.org/abs/2507.23726. arXiv:2507.23726 [cs].

Chen, X., Li, G., Wang, Z., Jin, B., Qian, C., Wang, Y., Wang, H., Zhang, Y., Zhang, D., Zhang, T., Tong, H., and Ji, H. RM-R1: Reward Modeling as Reasoning, May 2025b. URL http://arxiv.org/abs/2505.02387. arXiv:2505.02387 [cs].

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., Marris, L., Petulla, S., Gaffney, C., Aharoni, A., Lintz, N., Pais, T. C., Jacobsson, H., Szpektor, I., Jiang, N.-J., Haridasan, K., Omran, A., Saunshi, N., Bahri, D., Mishra, G., Chu, E., Boyd, T., Hekman, B., Parisi, A., Zhang, C., Kawintiranon, K., Bedrax-Weiss, T., Wang, O., Xu, Y., Purkiss, O., Mendlovic, U., Deutel, I., Nguyen, N., Langley, A., Korn, F., Rossazza, L., Ram, A., Waghmare, S., Miller, H., Byrd, N., Sheshan, A., Hadsell, R., Bhardwaj, S., Janus, P., Rissa, T., Horgan, D., Abdagic, A., Belenki, L., Allingham, J., Singh, A., Guidroz, T., Srinivasan, S., Schmit, H., Chiafullo, K., Elisseeff, A., Jha, N., Kolhar, P., Berrada, L., Ding, F., Si, X., Mallick, S. B., Och, F., Erell, S., Ni, E., Latkar, T., Yang, S., Sirkovic, P., Feng, Z., Leland, R., Hornung, R., Wu, G., Blundell, C., Alvari, H., Huang, P.-S., Yip, C., Deur, S., Liu, L., Surita, G., Duque, P., Damen, D., Jia, J., Guez, A., Mircea, M., Sinha, A., Magni, A., Stradomski, P., Marian, T., Gali, V., Chen, W., Husain, H., Singhal, A., Grewe, D., Aubet, F.-X., Song, S., Blanco, L., Rechis, L., Ho, L., Munoz, R., Zheng, K., Hamrick, J., Mather, K., Taitelbaum, H., Rutherford, E., Lei, Y., Chen, K., Shukla, A., Moreira, E., Doi, E., Isik, B., Shabat, N., Rogoziska, D., Kolipaka, K., Chang, J., Vuak, E., Venkatachary, S., Noghabi, S., Bharti, T., Jun, Y., Zaks, A., Green, S., Challagundla, J., Wong, W., Mohammad, M., Hirsch, D., Cheng, Y., Naim, I., Proleev, L., Vincent, D., Singh, A., Krikun, M., Krishnan, D., Ghahramani, Z., Atias, A., Aggarwal, R., Kirov, C., Vytiniotis, D., Koh, C., Chronopoulou, A., Dogra, P., Ion, V.-D., Tyen, G., Lee, J., Weissenberger, F., Strohman, T., Balakrishna, A., Rae, J., Velic, M., Liedekerke, R. d., Elyada, O., Yuan, W., Liu, C., Shani, L., Kishchenko, S., Alessio, B., Li, Y., Song, R., Kwei, S., Jankowski, O., Pappu, A., Namiki, Y., Ma, Y., Tripuraneni, N., Cherry, C., Ikonomidis, M., Ling, Y.-C., Ji, C., Westberg, B., Wright, A., Yu, D., Parkinson, D., Ramaswamy, S., Connor, J., Yeganeh, S. H., Grover, S., Kenwright, G., Litchev, L., Apps, C., Tomala, A., Halim, F., Castro-Ros, A., Li, Z., Boral, A., Sho, P., Yarom, M., Malmi, E., Klinghoffer, D., Lin, R., Ansell, A., S, P. K., Zhao, S., Zuo, S., Santoro, A., Cheng, H.-T., Demmessie, S., Liu, Y., Brichtova, N., Culp, A., Braun, N., Graur, D., Ng, W., Mehta, N., Phillips, A., Sundberg, P., Godbole, V., Liu, F., Katariya, Y., Rim, D., Seyedhosseini, M., Ammirati, S., Valfridsson, J., Malihi, M., Knight, T., Toor, A., Lampe, T., Ittycheriah, A., Chiang, L., Yeung, C., Frchette, A., Rao, J., Wang, H., Srivastava, H., Zhang, R., Rhodes, R., Brand, A., Weesner, D., Figotin, I., Gimeno, F., Fellinger, R., Marcenac, P., Leal, J., Marcus, E., Cotruta, V., Cabrera, R., Luo, S., Garrette, D., Axelrod, V., Baltateanu, S., Barker, D., Chen, D., Toma, H., Ingram, B., Riesa, J., Kulkarni, C., Zhang, Y., Liu, H., Wang, C., Polacek, M., Wu, W., Hui, K., Reyes, A. N., Su, Y., Barnes, M., Malhi, I., Siddiqui, A., Feng, Q., Damaschin, M., Pighin, D., Steiner, A., Yang, S., Boppana, R. S., Ivanov, S., Kandoor, A., Shah, A., Mujika, A., Huang, D., Choquette-Choo, C. A., Patel, M., Yu, T., Creswell, T., Jerry, Liu, Barros, C., Razeghi, Y., Roy, A., Culliton, P., Xiong, B., Pan, J., Strohmann, T., Powell, T., Seal, B., DeCarlo, D., Shyam, P., Katircioglu, K., Wang, X., Hardin, C., Odisho, I., Broder, J., Chang, O., Nair, A., Shtefan, A., O'Brien, M., Agarwal, M., Potluri, S., Goyal, S., Jhindal, A., Thakur, S., Stuken, Y., Lyon, J., Toutanova, K., Feng, F., Wu, A., Horn, B., Wang, A., Cullum, A., Taubman, G., Shrivastava, D., Shi, C., Tomlinson, H., Patel, R., Tu, T., Oflazer, A. M., Pongetti, F., Yang, M., Taga, A. A., Perot, V., Pierse, N. W., Han, F., Drori, Y., Iturrate, I., Chakrabarti, A., Yeung, L., Dopson, D., Chen, Y.-t., Kulshreshtha, A., Guo, T., Pham, P., Schuster, T., Chen, J., Polozov, A., Xing, J., Zhou, H., Kacham, P., Kukliansky, D., Miech, A., Yaroshenko, S., Chi, E., Douglas, S., Fei, H., Blondel, M., Myla, P., Madmoni, L., Wu, X., Keysers, D., Kjems, K., Albuquerque, I., Yu, L., D'sa, J., Plantan, M., Ionescu, V., Elias, J. S., Gupta, A., Vuyyuru, M. R., Alcober, F., Zhou, T., Ji, K., Hartmann, F., Puttagunta, S., Song, H., Amid, E., Stefanoiu, A., Lee, A., Pucciarelli, P., Wang, E., Raul, A., Petrov, S., Tian, I., Anklin, V., Nti, N., Gomes, V., Schumacher, M., Vesom, G., Panagopoulos, A., Bousmalis, K., Andor, D., Jacob, J., Zhang, Y., Rosgen, B., Kecman, M., Tung, M., Belias, A., Goodman, N., Covington, P., Wieder, B., Saxena, N., Davoodi, E., Huang, M., Maddineni, S., Roulet, V., Campbell-Ajala, F., Sessa, P. G., Xintian, Wu, Lai, G., Collins, P., Haig, A., Sakenas, V., Xu, X., Giustina, M., Shafey, L. E., Charoenpanit, P., Garg, S., Ainslie, J., Severson, B., Arenas, M. G., Pathak, S., Rajayogam, S., Feng, J., Bakker, M., Li, S., Wichers, N., Rogers, J., Geng, X., Li, Y., Jagerman, R., Jia, C., Olmert, N., Sharon, D., Mauger, M., Mariserla, S., Ma, H., Mohabey, M., Kim, K., Andreev, A., Pollom, S.,

Love, J., Jain, V., Agrawal, P., Schroecker, Y., Fortin, A., Warmuth, M., Liu, J., Leach, A., Blok, I., Girirajan, G. P., Aharoni, R., Uria, B., Sozanschi, A., Goldberg, D., Ionita, L., Ribeiro, M. T., Zlocha, M., Birodkar, V., Lachgar, S., Yuan, L., Choudhury, H., Ginsberg, M., Zheng, F., Dibb, G., Graves, E., Lokhande, S., Rasskin, G., Muraru, G.-C., Quick, C., Tata, S., Sermanet, P., Chawla, A., Karo, I., Wang, Y., Zhang, S., Keller, O., Dragan, A., Su, G., Chou, I., Liu, X., Tao, Y., Prabhakara, S., Wilson, M., Liu, R., Wang, S., Evans, G., Du, D., Castao, A., Prasad, G., Mahdy, M. E., Gerlach, S., Reid, M., Kahn, J., Zait, A., Pillai, T. S., Ulrich, T., Wang, G., Wassenberg, J., Farkash, E., Yalasangi, K., Wang, C., Bauza, M., Bucher, S., Liu, T., Yan, J., Leung, G., Sindhwani, V., Barnes, P., Singh, A., Jurin, I., Chang, J., Bhumihar, N. K., Eiger, S., Citovsky, G., Withbroe, B., Li, Z., Xue, S., Santo, N. D., Stoyanov, G., Raimond, Y., Zheng, S., Gao, Y., Listk, V., Kwasiborski, S., Saputro, R., Ozturel, A., Mallya, G., Majmundar, K., West, R., Caron, P., Wei, J., Castrejon, L., Vikram, S., Ramachandran, D., Dhawan, N., Park, J., Smoot, S., Driessche, G. v. d., Blau, Y., Malik, C., Liang, W., Hirsch, R., Santos, C. N. d., Weinstein, E., Oord, A. v. d., Lall, S., FitzGerald, N., Jiang, Z., Yang, X., Webster, D., Elqursh, A., Pope, A., Rotival, G., Raposo, D., Zhu, W., Dean, J., Alabed, S., Tran, D., Gupta, A., Gleicher, Z., Austin, J., Rosseel, E., Umekar, M., Das, D., Sun, Y., Chen, K., Misiunas, K., Zhou, X., Di, Y., Loo, A., Newlan, J., Li, B., Ramasesh, V., Xu, Y., Chen, A., Gandhe, S., Soricut, R., Gupta, N., Hu, S., El-Sayed, S., Garcia, X., Brusilovsky, I., Chen, P.-C., Bolt, A., Huang, L., Gurney, A., Zhang, Z., Pritzel, A., Wilkiewicz, J., Seybold, B., Shamanna, B. K., Fischer, F., Dean, J., Gill, K., Mcilroy, R., Bhowmick, A., Selier, J., Yang, A., Cheng, D., Magay, V., Tan, J., Varma, D., Walder, C., Kocisky, T., Nakashima, R., Natsev, P., Kwong, M., Gog, I., Zhang, C., Dieleman, S., Jimma, T., Ryabtsev, A., Brahma, S., Steiner, D., Du, D., uul, A., ani, M., Raghavachari, M., Gierke, W., Zheng, Z., Petrova, D., Dauphin, Y., Liu, Y., Kessler, I., Hand, S., Duvarney, C., Kim, S., Lee, H., Hussenot, L., Hui, J., Smith, J., Jain, D., Xia, J., Tomar, G. S., Amiri, K., Phan, D., Fuchs, F., Weyand, T., Tomasev, N., Cordell, A., Liu, X., Mallinson, J., Joshi, P., Crawford, A., Suggala, A., Chien, S., Fernando, N., Sanchez-Vargas, M., Williams, D., Crone, P., Luo, X., Karpov, I., Shan, J., Thurk, T., Strudel, R., Voigtlaender, P., Patil, P., Dozat, T., Khodaei, A., Singla, S., Ambroszczyk, P., Wu, Q., Chang, Y., Roark, B., Hegde, C., Ding, T., Filos, A., Wu, Z., Pinto, A. S., Liu, S., Khanna, S., Pandey, A., Mcloughlin, S., Li, Q., Haves, S., Zhou, A., Buchatskaya, E., Leal, I., Boursac, P. d., Akazawa, N., Anderson, N., Chen, T., Somandepalli, K., Liang, C., Goenka, S., Winkler, S., Grushetsky, A., Ding, Y., Smith, J., Ye, F., Pont-Tuset, J., Li, E., Li, R., Golany, T., Wegner, D., Jiang, T., Barak, O., Shangguan, Y., Vrtes, E., Wong, R., Bornschein, J., Tudor, A., Bevilacqua, M., Schaul, T., Rawat, A. S., Zhao, Y., Axiotis, K., Meng, L., McLean, C., Lai, J., Beattie, J., Kushman, N., Liu, Y., Kutzman, B., Lang, F., Ye, J., Netrapalli, P., Mishra, P., Khan, M., Goel, M., Willoughby, R., Tian, D., Zhuang, H., Chen, J. D., Tsai, Z., Kementsietsidis, T., Khare, A., Keeling, J., Xu, K., Waters, N., Altch, F., Popat, A., Mittal, B., Saxton, D., Badawy, D. E., Mathieu, M., Zheng, Z., Zhou, H., Ranka, N., Shin, R., Duan, Q., Salimans, T., Mihailescu, I., Shaham, U., Chang, M.-W., Assael, Y., Dikkala, N., Izzard, M., Cohen-Addad, V., Graves, C., Feinberg, V., Chung, G., Strouse, D. J., Karmon, D., Sharifzadeh, S., Ashwood, Z., Pham, K., Blanton, J., Vasiloff, A., Barber, J., Geller, M., Zhou, A., Zubach, F., Huang, T.-K., Zhang, L., Gupta, H., Young, M., Proskurnia, J., Votel, R., Gabeur, V., Barcik, G., Tripathi, A., Yu, H., Yan, G., Changpinyo, B., Paveti, F., Coyle, A., Fujii, Y., Mendez, J. G., Zhou, T., Rajamani, H., Hechtman, B., Cao, E., Juan, D.-C., Tan, Y.-X., Dalibard, V., Du, Y., Clay, N., Yao, K., Jia, W., Vijaykumar, D., Zhou, Y., Bai, X., Hung, W.-C., Pecht, S., Todorov, G., Khadke, N., Gupta, P., Lahoti, P., Autef, A., Duddu, K., Lee-Thorp, J., Bykovsky, A., Misiunas, T., Flennerhag, S., Thangaraj, S., McGiffin, J., Nado, Z., Kunesch, M., Noever, A., Hertz, A., Liang, M., Stone, V., Palmer, E., Daruki, S., Pramanik, A., Pder, S., Kyker, A., Khan, M., Sluzhaev, E., Ritter, M., Ruderman, A., Zhou, W., Nagpal, C., Vodrahalli, K., Necula, G., Barham, P., Pavlick, E., Hartford, J., Shafran, I., Zhao, L., Mikua, M., Eccles, T., Shimokawa, H., Garg, K., Vilnis, L., Chen, H., Shumailov, I., Lee, K.-H., Abdelhamed, A., Xie, M., Cohen, V., Hlavnova, E., Malkin, D., Sitawarin, C., Lottes, J., Coquinot, P., Yu, T., Kumar, S., Zhang, J., Mahendru, A., Ahmed, Z., Martens, J., Chen, T., Boag, A., Peng, D., Devin, C., Klimovskiy, A., Phuong, M., Vainstein, D., Xie, J., Ramabhadran, B., Howard, N., Yu, X., Goswami, G., Cui, J., Shleifer, S., Pinto, M., Yeh, C.-K., Yang, M.-H., Javanmardi, S., Ethier, D., Lee, C., Orbay, J., Kotecha, S., Bromberg, C., Shaw, P., Thornton, J., Rosenthal, A. G., Gu, S., Thomas, M., Gemp, I., Ayyar, A., Ushio, A., Selvan, A., Wee, J., Liu, C., Majzoubi, M., Yu, W., Abernethy, J., Liechty, T., Pan, R., Nguyen, H., Qiong, Hu, Perrin, S., Arora, A., Pitler, E., Wang, W., Shivakumar, K., Prost, F., Limonchik, B., Wang, J., Gao, Y., Cour, T., Buch, S., Gui, H., Ivanova, M., Neubeck, P., Chan, K., Kim, L., Chen, H., Goyal, N., Chung, D.-W., Liu, L., Su, Y., Petrushkina, A., Shen, J., Joulin, A., Xu, Y., Lin, S. X., Kulizhskaya, Y., Chelba, C., Vasudevan, S., Collins, E., Bashlovkina, V., Lu, T., Fritz, D., Park, J., Zhou, Y., Su, C., Tanburn, R., Sushkov, M., Rasquinha, M., Li, J., Prendki, J., Li, Y., LV, P., Sharma, S., Fitoussi, H., Huang, H., Dai, A., Dao, P., Burrows, M., Prior, H., Qin, D., Pundak, G., Sjoesund, L. L., Khurshudov, A., Zhu, Z., Webson, A., Kemp, E., Tan, T., Agrawal, S., Sargsyan, S., Cheng, L., Stephan, J.,

Kwiatkowski, T., Reid, D., Byravan, A., Michaely, A. H., Heess, N., Zhou, L., Goenka, S., Carpenter, V., Levskaya, A., Wang, B., Roberts, R., Leblond, R., Chikkerur, S., Ginzburg, S., Chang, M., Riachi, R., Chuqiao, Xu, Borsos, Z., Pliskin, M., Pawar, J., Lustman, M., Kirkwood, H., Anand, A., Chaudhary, A., Kalb, N., Milan, K., Augenstein, S., Goldie, A., Prince, L., Raman, K., Sun, Y., Xia, V., Cohen, A., Huo, Z., Camp, J., Ellis, S., Zilka, L., Torres, D. V., Patel, L., Arora, S., Chan, B., Adler, J., Ayoub, K., Liang, J., Jamil, F., Jiang, J., Baumgartner, S., Sun, H., Karov, Y., Akulov, Y., Zheng, H., Cai, I., Fantacci, C., Rubin, J., Acha, A. R., Wang, M., D'Souza, N., Sathyanarayana, R., Dai, S., Rowe, S., Simanovsky, A., Goldman, O., Kuang, Y., Pan, X., Rosenberg, A., Rojas-Esponda, T., Dutta, P., Zeng, A., Jurenka, I., Farquhar, G., Bansal, Y., Iqbal, S., Roelofs, B., Joung, G.-Y., Beak, P., Ryu, C., Poplin, R., Wu, Y., Alayrac, J.-B., Buthpitiya, S., Ronneberger, O., Habtegebriel, C., Li, W., Cavallaro, P., Wei, A., Bensky, G., Denk, T., Ganapathy, H., Stanway, J., Joshi, P., Bertolini, F., Lo, J., Ma, O., Charles, Z., Sampemane, G., Sahni, H., Chen, X., Askham, H., Gaddy, D., Young, P., Tan, J., Eyal, M., Brainskas, A., Zhong, L., Wu, Z., Epstein, M., Bailey, K., Hard, A., Lee, K., Goldshtein, S., Ruiz, A., Badawi, M., Lochbrunner, M., Kearns, J. K., Brown, A., Pardo, F., Weber, T., Yang, H., Jiang, P.-P., Akin, B., Fu, Z., Wainwright, M., Zou, C., Gaba, M., Manzagol, P.-A., Kan, W., Song, Y., Zainullina, K., Lin, R., Ko, J., Deshmukh, S., Jindal, A., Svensson, J., Tyam, D., Zhao, H., Kaeser-Chen, C., Baird, S., Moradi, P., Hall, J., Guo, Q., Tsang, V., Liang, B., Pereira, F., Ganesh, S., Korotkov, I., Adamek, J., Thiagarajan, S., Tran, V., Chen, C., Tar, C., Jain, S., Dasgupta, I., Bilal, T., Reitter, D., Zhao, K., Vezzani, G., Gehman, Y., Mehta, P., Beltrone, L., Dotiwalla, X., Guadarrama, S., Abbas, Z., Karp, S., Georgiev, P., Ferng, C.-S., Brockschmidt, M., Peng, L., Hirnschall, C., Verma, V., Bi, Y., Xiao, Y., Dabush, A., Xu, K., Wallis, P., Parker, R., Wang, Q., Xu, Y., Safarli, I., Tewari, D., Zhang, Y., Kim, S., Gesmundo, A., Thomas, M., Levi, S., Chowdhury, A., Rao, K., Garst, P., Conway-Rahman, S., Ran, H., McKinney, K., Xiao, Z., Yu, W., Agrawal, R., Stjerngren, A., Ionescu, C., Chen, J., Sharma, V., Chiu, J., Liu, F., Franko, K., Sanford, C., Cai, X., Michel, P., Ganapathy, S., Labanowski, J., Garrett, Z., Vargas, B., Sun, S., Gale, B., Buschmann, T., Desjardins, G., Ghelani, N., Jain, P., Verma, M., Asawaroengchai, C., Eisenschlos, J., Harlalka, J., Kazawa, H., Metzler, D., Howland, J., Jian, Y., Ades, J., Shah, V., Gangwani, T., Lee, S., Ring, R., Hernandez, S. M., Reich, D., Sinha, A., Sathe, A., Kovac, J., Gill, A., Kannan, A., D'olimpio, A., Sevenich, M., Whang, J., Kim, B., Sim, K. C., Chen, J., Zhang, J., Lall, S., Matias, Y., Jia, B., Friesen, A., Nasso, S., Thapliyal, A., Perozzi, B., Yu, T., Shekhawat, A., Huda, S., Grabowski, P., Wang, E., Sreevatsa, A., Dib, H., Hassen, M., Schuh, P., Milutinovic, V., Welty, C., Quinn, M., Shah, A., Wang, B., Barth-Maron, G., Frye, J., Axelsson, N., Zhu, T., Ma, Y., Giannoumis, I., Sedghi, H., Ye, C., Luan, Y., Aydin, K., Chandra, B., Sampathkumar, V., Huang, R., Lavrenko, V., Eleryan, A., Hong, Z., Hansen, S., Carthy, S. M., Samanta, B., evid, D., Wang, X., Li, F., Voznesensky, M., Hoffman, M., Terzis, A., Sehwag, V., Fidel, G., He, L., Cai, M., He, Y., Feng, A., Nikoltchev, M., Phatale, S., Chase, J., Lawton, R., Zhang, M., Ouyang, T., Tragut, M., Manshadi, M. H., Narayanan, A., Shen, J., Gao, X., Bolukbasi, T., Roy, N., Li, X., Golovin, D., Panait, L., Qin, Z., Han, G., Anthony, T., Kudugunta, S., Patraucean, V., Ray, A., Chen, X., Yang, X., Bhatia, T., Talluri, P., Morris, A., Ranatovi, A., Brownfield, B., An, J., Peng, S., Kane, P., Zheng, C., Duduta, N., Kessinger, J., Noraky, J., Liu, S., Rong, K., Velikovi, P., Rush, K., Goldin, A., Wei, F., Garlapati, S. M. R., Pantofaru, C., Kwon, O., Ni, J., Noland, E., Trapani, J. D., Beaufays, F., Roy, A. G., Chow, Y., Turker, A., Cideron, G., Mei, L., Clark, J., Dou, Q., Bonjak, M., Leith, R., Du, Y., Yazdanbakhsh, A., Nasr, M., Kwak, C., Sheth, S. S., Kaskasoli, A., Anand, A., Lakshminarayanan, B., Jerome, S., Bieber, D., Chu, C.-T., Senges, A., Shen, T., Sridhar, M., Ndebele, N., Beyret, B., Mohamed, S., Chen, M., Freitag, M., Guo, J., Liu, L., Roit, P., Chen, H., Yan, S., Stone, T., Co-Reyes, J. D., Cole, J., Scellato, S., Azizi, S., Hashemi, H., Jin, A., Iyer, A., Valentine, M., Gyrgy, A., Ahuja, A., Diaz, D. H., Lee, C.-Y., Clement, N., Kong, W., Garmon, D., Watts, I., Bhatia, K., Gupta, K., Miecnikowski, M., Vallet, H., Taly, A., Loper, E., Joshi, S., Atwood, J., Chick, J., Collier, M., Iliopoulos, F., Trostle, R., Gunel, B., Leal-Cavazos, R., Hrafnkelsson, A. M., Guzman, M., Ju, X., Forbes, A., Emond, J., Chauhan, K., Caine, B., Xiao, L., Zeng, W., Moufarek, A., Murphy, D., Meng, M., Gupta, N., Riedel, F., Das, A., Lawal, E., Narayan, S., Sosea, T., Swirhun, J., Friso, L., Neyshabur, B., Lu, J., Girgin, S., Wunder, M., Yvinec, E., Pyne, A., Carbune, V., Rijhwani, S., Guo, Y., Doshi, T., Briukhov, A., Bain, M., Hitron, A., Wang, X., Gupta, A., Chen, K., Du, C., Zhang, W., Shah, D., Akula, A., Dylla, M., Kachra, A., Kuo, W., Zou, T., Wang, L., Xu, L., Zhu, J., Snyder, J., Menon, S., Firat, O., Mordatch, I., Yuan, Y., Ponomareva, N., Blevins, R., Moore, L., Wang, W., Chen, P., Scholz, M., Dwornik, A., Lin, J., Li, S., Antognini, D., I, T., Song, X., Miller, M., Kalra, U., Raveret, A., Akerlund, O., Wu, F., Nystrom, A., Godbole, N., Liu, T., DeBalsi, H., Zhao, J., Liu, B., Caciularu, A., Lax, L., Khandelwal, U., Langston, V., Bailey, E., Lattanzi, S., Wang, Y., Kovelamudi, N., Mondal, S., Guruganesh, G., Hua, N., Roval, O., Wesoowski, P., Ingale, R., Halcrow, J., Sohn, T., Angermueller, C., Raad, B., Stickgold, E., Lu, E., Kosik, A., Xie, J., Lillicrap, T., Huang, A., Zhang, L. L., Paulus, D., Farabet, C., Wertheim, A., Wang, B., Joshi, R., Ko, C.-l., Wu, Y., Agrawal, S., Lin, L., Sheng, X., Sung, P., Breland-King, T., Butterfield, C., Gawde,

S., Singh, S., Zhang, Q., Apte, R., Shetty, S., Hutter, A., Li, T., Salesky, E., Lebron, F., Kanerva, J., Paganini, M., Nguyen, A., Vallu, R., Peter, J.-T., Velury, S., Kao, D., Hoover, J., Bortsova, A., Bishop, C., Jakobovits, S., Agostini, A., Agarwal, A., Liu, C., Kwong, C., Tavakkol, S., Bica, I., Greve, A., GP, A., Marcus, J., Hou, L., Duerig, T., Moroshko, R., Lacey, D., Davis, A., Amelot, J., Wang, G., Kim, F., Strinopoulos, T., Wan, H., Lan, C. L., Krishnan, S., Tang, H., Humphreys, P., Bai, J., Shtacher, I. H., Machado, D., Pang, C., Burke, K., Liu, D., Aravamudhan, R., Song, Y., Hirst, E., Singh, A., Jou, B., Bai, L., Piccinno, F., Fu, C. K., Alazard, R., Meiri, B., Winter, D., Chen, C., Zhang, M., Heitkaemper, J., Lambert, J., Lee, J., Frmmgen, A., Rogulenko, S., Nair, P., Niemczyk, P., Bulyenov, A., Xu, B., Shemtov, H., Zadimoghaddam, M., Toropov, S., Wirth, M., Dai, H., Gollapudi, S., Zheng, D., Kurakin, A., Lee, C., Bullard, K., Serrano, N., Balazevic, I., Li, Y., Schalkwyk, J., Murphy, M., Zhang, M., Sequeira, K., Datta, R., Agrawal, N., Sutton, C., Attaluri, N., Chiang, M., Farhan, W., Thornton, G., Lin, K., Choma, T., Nguyen, H., Dasgupta, K., Robinson, D., Coma, I., Riley, M., Pillai, A., Mustafa, B., Golan, B., Zandieh, A., Lespiau, J.-B., Porter, B., Ross, D., Rajayogam, S., Agarwal, M., Venugopalan, S., Shahriari, B., Yan, Q., Xu, H., Tobin, T., Dubov, P., Shi, H., Recasens, A., Kovsharov, A., Borgeaud, S., Dery, L., Vasanth, S., Gribovskaya, E., Qiu, L., Mahdieh, M., Skut, W., Nielsen, E., Zheng, C. J., Yu, A., Bostock, C. G., Gupta, S., Archer, A., Rawles, C., Davies, E., Svyatkovskiy, A., Tsai, T., Halpern, Y., Reisswig, C., Wydrowski, B., Chang, B., Puigcerver, J., Taege, M. H., Li, J., Schnider, E., Li, X., Dena, D., Xu, Y., Telang, U., Shi, T., Zen, H., Kastner, K., Ko, Y., Subramaniam, N., Kumar, A., Blois, P., Dai, Z., Wieting, J., Lu, Y., Zeldes, Y., Xie, T., Hauth, A., ifrea, A., Li, Y., El-Husseini, S., Abolafia, D., Zhou, H., Ding, W., Ghalebikesabi, S., Gua, C., Maksai, A., Weisz, ., Arik, S., Sukhanov, N., wietlik, A., Jia, X., Yu, L., Wang, W., Brand, M., Bloxwich, D., Kirmani, S., Chen, Z., Go, A., Sprechmann, P., Kannen, N., Carin, A., Sandhu, P., Edkins, I., Nooteboom, L., Gupta, J., Maggiore, L., Azizi, J., Pritch, Y., Yin, P., Gupta, M., Tarlow, D., Smith, D., Ivanov, D., Babaeizadeh, M., Goel, A., Kambala, S., Chu, G., Kastelic, M., Liu, M., Soltau, H., Stone, A., Agrawal, S., Kim, M., Soparkar, K., Tadepalli, S., Bunyan, O., Soh, R., Kannan, A., Kim, D. Y., Chen, B. J., Halumi, A., Roy, S., Wang, Y., Sercinoglu, O., Gibson, G., Bhatnagar, S., Sano, M., Dincklage, D. v., Ren, Q., Mitrevski, B., Olk, M., She, J., Doersch, C., Jilei, Wang, Liu, B., Tan, Q., Yakar, T., Warkentin, T., Ramirez, A., Lebsack, C., Dillon, J., Mathews, R., Cobley, T., Wu, Z., Chen, Z., Simon, J., Nath, S., Sainath, T., Bendebury, A., Julian, R., Mankalale, B., urko, D., Zacchello, P., Brown, A. R., Sodhia, K., Howard, H., Caelles, S., Gupta, A., Evans, G., Bulanova, A., Katzen, L., Goldenberg, R., Tsitsulin, A., Stanton, J., Schillings, B., Kovalev, V., Fry, C., Shah, R., Lin, K., Upadhyay, S., Li, C., Radpour, S., Maggioni, M., Xiong, J., Haas, L., Brennan, J., Kamath, A., Savinov, N., Nagrani, A., Yacovone, T., Kappedal, R., Andriopoulos, K., Lao, L., Li, Y., Rozhdestvenskiy, G., Hashimoto, K., Audibert, A., Austin, S., Rodriguez, D., Ruoss, A., Honke, G., Karkhanis, D., Xiong, X., Wei, Q., Huang, J., Leng, Z., Premachandran, V., Bileschi, S., Evangelopoulos, G., Mensink, T., Pavagadhi, J., Teplyashin, D., Chang, P., Xue, L., Tanzer, G., Goldman, S., Patel, K., Li, S., Wiesner, J., Zheng, I., Stewart-Binks, I., Han, J., Li, Z., Luo, L., Lenc, K., Lui, M., Xue, F., Mullins, R., Guseynov, A., Chang, C.-C., Galatzer-Levy, I., Zhang, A., Bingham, G., Hu, G., Hartman, A., Ma, Y., Griffith, J., Irpan, A., Radebaugh, C., Yue, S., Fan, L., Ungureanu, V., Sorokin, C., Teufel, H., Li, P., Anil, R., Paparas, D., Wang, T., Lin, C.-C., Peng, H., Shum, M., Petrovic, G., Brady, D., Nguyen, R., Macherey, K., Li, Z., Singh, H., Yenugula, M., Iinuma, M., Chen, X., Kopparapu, K., Stern, A., Dave, S., Thekkath, C., Perot, F., Kumar, A., Li, F., Xiao, Y., Bilotti, M., Bateni, M. H., Noble, I., Lee, L., Vzquez-Reina, A., Salazar, J., Yang, X., Wang, B., Gruzewska, E., Rao, A., Raghuram, S., Xu, Z., Ben-David, E., Mei, J., Dalmia, S., Zhang, Z., Liu, Y., Bansal, G., Pankov, H., Schwarcz, S., Burns, A., Chan, C., Sanghai, S., Liang, R., Liang, E., He, A., Stuart, A., Narayanan, A., Zhu, Y., Frank, C., Fatemi, B., Sabne, A., Lang, O., Bhattacharya, I., Settle, S., Wang, M., McMahan, B., Tacchetti, A., Soares, L. B., Hadian, M., Cabi, S., Chung, T., Putikhin, N., Li, G., Chen, J., Tarango, A., Michalewski, H., Kazemi, M., Masoom, H., Sheftel, H., Shivanna, R., Vadali, A., Comanescu, R., Reid, D., Moore, J., Neelakantan, A., Sander, M., Herzig, J., Rosenberg, A., Dehghani, M., Choi, J. D., Fink, M., Hayes, R., Ge, E., Weng, S., Ho, C.-H., Karro, J., Krishna, K., Thiet, L. N., Skerry-Ryan, A., Eppens, D., Andreetto, M., Sarma, N., Bonacina, S., Ayan, B. K., Nawhal, M., Shan, Z., Dusenberry, M., Thakoor, S., Gubbi, S., Nguyen, D. D., Tsarfaty, R., Albanie, S., Mitrovi, J., Gandhi, M., Chen, B.-J., Epasto, A., Stephanov, G., Jin, Y., Gehman, S., Amini, A., Weber, J., Behbahani, F., Xu, S., Allamanis, M., Chen, X., Ott, M., Sha, C., Jastrzebski, M., Qi, H., Greene, D., Wu, X., Toki, A., Vlasic, D., Shapiro, J., Kotikalapudi, R., Shen, Z., Saeki, T., Xie, S., Cassirer, A., Bharadwaj, S., Kiyono, T., Bhojanapalli, S., Rosenfeld, E., Ritter, S., Mao, J., Oliveira, J. G., Egyed, Z., Bandemer, B., Parisotto, E., Kinoshita, K., Pluto, J., Maniatis, P., Li, S., Guo, Y., Ghiasi, G., Tarbouriech, J., Chatterjee, S., Jin, J., Katrina, Xu, Palomaki, J., Arnold, S., Sewak, M., Piccinini, F., Sharma, M., Albrecht, B., Purser-haskell, S., Vaswani, A., Chen, C., Wisniewski, M., Cao, Q., Aslanides, J., Phu, N. M., Sieb, M., Agubuzu, L., Zheng, A., Sohn, D., Selvi, M., Andreassen, A., Subudhi, K., Eruvbetine, P., Woodman, O., Mery, T., Krause, S., Ren, X., Ma, X., Luo, J., Chen,

D., Fan, W., Griffiths, H., Schuler, C., Li, A., Zhang, S., Sarr, J.-M., Luo, S., Patana, R., Watson, M., Naboulsi, D., Collins, M., Sidhwani, S., Hoogeboom, E., Silver, S., Caveness, E., Zhao, X., Rodriguez, M., Deines, M., Bai, L., Griffin, P., Tagliasacchi, M., Xue, E., Babbula, S. R., Pang, B., Ding, N., Shen, G., Peake, E., Crocker, R., Raghvendra, S. S., Swisher, D., Han, W., Singh, R., Wu, L., Pchelin, V., Munkhdalai, T., Alon, D., Bacon, G., Robles, E., Bulian, J., Johnson, M., Powell, G., Ferreira, F. T., Li, Y., Benzing, F., Velimirovi, M., Soyer, H., Kong, W., Tony, Nguyłn, Yang, Z., Liu, J., Amersfoort, J. v., Gillick, D., Sun, B., Rauschmayr, N., Zhang, K., Zhan, S., Zhou, T., Frolov, A., Yang, C., Vnukov, D., Rouillard, L., Li, H., Mandhane, A., Fallen, N., Venkataraman, R., Hu, C. H., Brennan, J., Lee, J., Chang, J., Sundermeyer, M., Pan, Z., Ke, R., Tong, S., Fabrikant, A., Bono, W., Gu, J., Foley, R., Mao, Y., Delakis, M., Bhaswar, D., Frostig, R., Li, N., Zipori, A., Hope, C., Kozlova, O., Mishra, S., Djolonga, J., Schiff, C., Merey, M. A., Briakou, E., Morgan, P., Wan, A., Hassidim, A., Skerry-Ryan, R. J., Sengupta, K., Jasarevic, M., Kallakuri, P., Kunkle, P., Brennan, H., Lieber, T., Mansoor, H., Walker, J., Zhang, B., Xie, A., ui, G., Chukwuka, A., Druinsky, A., Cho, D., Yao, R., Naeem, F., Butt, S., Kim, E., Jia, Z., Jordan, M., Lelkes, A., Kurzeja, M., Wang, S., Zhao, J., Over, A., Chakladar, A., Prasetya, M., Jha, N., Ganapathy, S., Cong, Y., Shroff, P., Saroufim, C., Miryoosefi, S., Hammad, M., Nasir, T., Xi, W., Gao, Y., Maeng, Y., Hora, B., Cheng, C.-Y., Haghani, P., Lewenberg, Y., Lu, C., Matysiak, M., Raisinghani, N., Wang, H., Baugher, L., Sukthankar, R., Giang, M., Schultz, J., Fiedel, N., Chen, M., Lee, C.-C., Dey, T., Zheng, H., Paul, S., Smith, C., Ly, A., Wang, Y., Bansal, R., Perz, B., Ricco, S., Blank, S., Keshava, V., Sharma, D., Chow, M., Lad, K., Jalan, K., Osindero, S., Swanson, C., Scott, J., Ili, A., Li, X., Jonnalagadda, S. R., Soudagar, A. S., Xiong, Y., Batsaikhan, B.-O., Jarrett, D., Kumar, N., Shah, M., Lawlor, M., Waters, A., Graham, M., May, R., Ramos, S., Lefdal, S., Cankara, Z., Cano, N., O'Donoghue, B., Borovik, J., Liu, F., Grimstad, J., Alnahlawi, M., Tsihlas, K., Hudson, T., Grigorev, N., Jia, Y., Huang, T., Igwe, T. P., Lebedev, S., Tang, X., Krivokon, I., Garcia, F., Tan, M., Jia, E., Stys, P., Vashishth, S., Liang, Y., Venkatraman, B., Gu, C., Kementsietsidis, A., Zhu, C., Jung, J., Bai, Y., Hosseini, M. J., Ahmed, F., Gupta, A., Yuan, X., Ashraf, S., Nigam, S., Vasudevan, G., Awasthi, P., Gilady, A. M., Mariet, Z., Eskander, R., Li, H., Hu, H., Garrido, G., Schlattner, P., Zhang, G., Saxena, R., Devi, P., Muralidharan, K., Murthy, A., Zhou, Y., Choi, M., Wongpanich, A., Wang, Z., Shah, P., Xu, Y., Huang, Y., Spencer, S., Chen, A., Cohan, J., Wang, J., Tompson, J., Wu, J., Haroun, R., Li, H., Huergo, B., Yang, F., Yin, T., Wendt, J., Bendersky, M., Chaabouni, R., Snaider, J., Ferret, J., Jindal, A., Thompson, T., Xue, A., Bishop, W., Phal, S. M., Sharma, A., Sung, Y., Radhakrishnan, P., Shomrat, M., Ingle, R., Vij, R., Gilmer, J., Istin, M. D., Sobell, S., Lu, Y., Nottage, E., Sadigh, D., Willcock, J., Zhang, T., Xu, S., Brown, S., Lee, K., Wang, G., Zhu, Y., Tay, Y., Kim, C., Gutierrez, A., Sharma, A., Xian, Y., Seo, S., Cui, C., Pochernina, E., Baetu, C., Jastrzbski, K., Ly, M., Elhawaty, M., Suh, D., Sezener, E., Wang, P., Yuen, N., Tucker, G., Cai, J., Yang, Z., Wang, C., Muzio, A., Qian, H., Yoo, J., Lockhart, D., McKee, K. R., Guo, M., Mehrotra, M., Mendona, A., Mehta, S. V., Ben, S., Tekur, C., Mu, J., Zhu, M., Krakovna, V., Lee, H., Maschinot, A. J., Cevey, S., Choe, H., Bai, A., Srinivasan, H., Gasaway, D., Young, N., Siegler, P., Holtmann-Rice, D., Piratla, V., Baumli, K., Yogev, R., Hofer, A., Hasselt, H. v., Grant, S., Chervonyi, Y., Silver, D., Hogue, A., Agarwal, A., Wang, K., Singh, P., Flynn, F., Lipschultz, J., David, R., Bellot, L., Yang, Y.-Y., Le, L., Graziano, F., Olszewska, K., Hui, K., Maurya, A., Parotsidis, N., Chen, W., Oguntebi, T., Kelley, J., Baddepudi, A., Mauerer, J., Shaw, G., Siegman, A., Yang, L., Shetty, S., Roy, S., Song, Y., Stokowiec, W., Burnell, R., Savant, O., Busa-Fekete, R., Miao, J., Ghosh, S., MacDermed, L., Lippe, P., Dektiarev, M., Behrman, Z., Mentzer, F., Nguyen, K., Wei, M., Verma, S., Knutsen, C., Dasari, S., Yan, Z., Mitrichev, P., Wang, X., Shejwalkar, V., Austin, J., Sunkara, S., Potti, N., Virin, Y., Wright, C., Liu, G., Riva, O., Pot, E., Kochanski, G., Le, Q., Balasubramaniam, G., Dhar, A., Liao, Y., Bloniarz, A., Shukla, D., Cole, E., Lee, J., Zhang, S., Kafle, S., Vashishtha, S., Mahmoudieh, P., Chen, G., Hoffmann, R., Srinivasan, P., Lago, A. D., Shalom, Y. B., Wang, Z., Elabd, M., Sharma, A., Oh, J., Kothawade, S., Le, M., Monteiro, M., Yang, S., Alarakyia, K., Geirhos, R., Mincu, D., Garnes, H., Kobayashi, H., Mariooryad, S., Krasowiak, K., Zhixin, Lai, Mourad, S., Wang, M., Bu, F., Aharoni, O., Chen, G., Goyal, A., Zubov, V., Bapna, A., Dabir, E., Kothari, N., Lamerigts, K., Cao, N. D., Shar, J., Yew, C., Kulkarni, N., Mahaarachchi, D., Joshi, M., Zhu, Z., Lichtarge, J., Zhou, Y., Muckenhirn, H., Selo, V., Vinyals, O., Chen, P., Brohan, A., Mehta, V., Cogan, S., Wang, R., Geri, T., Ko, W.-J., Chen, W., Viola, F., Shivam, K., Wang, L., Elish, M. C., Popa, R. A., Pereira, S., Liu, J., Koster, R., Kim, D., Zhang, G., Ebrahimi, S., Talukdar, P., Zheng, Y., Poklukar, P., Mikhalap, A., Johnson, D., Vijayakumar, A., Omernick, M., Dibb, M., Dubey, A., Hu, Q., Suman, A., Aggarwal, V., Kornakov, I., Xia, F., Lowe, W., Kolganov, A., Xiao, T., Nikolaev, V., Hemingray, S., Li, B., Iljazi, J., Rybiski, M., Sandhu, B., Lu, P., Luong, T., Jenatton, R., Govindaraj, V., Hui, Li, Dulac-Arnold, G., Park, W., Wang, H., Modi, A., Pouget-Abadie, J., Greller, K., Gupta, R., Berry, R., Ramachandran, P., Xie, J., McCafferty, L., Wang, J., Gupta, K., Lim, H., Bratani, B., Brock, A., Akolzin, I., Sproch, J., Karliner, D., Kim, D., Goedeckemeyer, A., Shazeer, N., Schmid, C., Calandriello, D., Bhatia, P., Choromanski, K., Montgomery, C.,

Dua, D., Ramalho, A., King, H., Gao, Y., Nguyen, L., Lindner, D., Pitta, D., Johnson, O., Salama, K., Ardila, D., Han, M., Farnese, E., Odoom, S., Wang, Z., Ding, X., Rink, N., Smith, R., Lehri, H. T., Cohen, E., Vats, N., He, T., Gopavarapu, P., Paszke, A., Patel, M., Gansbeke, W. V., Loher, L., Castro, L., Voitovich, M., Glehn, T. v., George, N., Niklaus, S., Eaton-Rosen, Z., Rakievi, N., Jue, E., Perel, S., Zhang, C., Bahat, Y., Pouget, A., Xing, Z., Huot, F., Shenoy, A., Bos, T., Coriou, V., Richter, B., Noy, N., Wang, Y., Ontanon, S., Qin, S., Makarchuk, G., Hassabis, D., Li, Z., Sharma, M., Venkatesan, K., Kemaev, I., Daniel, R., Huang, S., Shah, S., Ponce, O., Warren, Chen, Faruqui, M., Wu, J., Andai, S., Payrits, S., McDuff, D., Hume, T., Cao, Y., Tessler, M. H., Wang, Q., Wang, Y., Rendulic, I., Agustsson, E., Johnson, M., Lando, T., Howard, A., Padmanabhan, S. G. S., Daswani, M., Banino, A., Kilgore, M., Heek, J., Ji, Z., Caceres, A., Li, C., Kassner, N., Vlaskin, A., Liu, Z., Grills, A., Hou, Y., Sukkerd, R., Cheon, G., Shetty, N., Markeeva, L., Stanczyk, P., Iyer, T., Gong, Y., Gao, S., Gopalakrishnan, K., Blyth, T., Reynolds, M., Bhoopchand, A., Bilenko, M., Gharibian, D., Zayats, V., Faust, A., Singh, A., Ma, M., Jiao, H., Vijayanarasimhan, S., Aroyo, L., Yadav, V., Chakera, S., Kakarla, A., Meshram, V., Gregor, K., Botea, G., Senter, E., Jia, D., Kovacs, G., Sharma, N., Baur, S., Kang, K., He, Y., Zhuo, L., Kostelac, M., Laish, I., Peng, S., O'Bryan, L., Kasenberg, D., Rao, G. R., Leurent, E., Zhang, B., Stevens, S., Salazar, A., Zhang, Y., Lobov, I., Walker, J., Porter, A., Redshaw, M., Ke, H., Rao, A., Lee, A., Lam, H., Moffitt, M., Kim, J., Qiao, S., Koo, T., Dadashi, R., Song, X., Sundararajan, M., Xu, P., Kawamoto, C., Zhong, Y., Barbu, C., Reddy, A., Verzetti, M., Li, L., Papamakarios, G., Klimczak-Pluciska, H., Cassin, M., Kavukcuoglu, K., Swavely, R., Vaucher, A., Zhao, J., Hemsley, R., Tschannen, M., Ge, H., Menghani, G., Yu, Y., Ha, N., He, W., Wu, X., Song, M., Sterneck, R., Zinke, S., Calian, D. A., Marsden, A., Ruiz, A. C., Hessel, M., Gueta, A., Lee, B., Farris, B., Gupta, M., Li, Y., Saleh, M., Misra, V., Xiao, K., Mendolicchio, P., Buttimore, G., Krayvanova, V., Nayakanti, N., Wiethoff, M., Pande, Y., Mirhoseini, A., Lao, N., Liu, J., Hua, Y., Chen, A., Malkov, Y., Kalashnikov, D., Gupta, S., Audhkhasi, K., Zhai, Y., Kopalle, S., Jain, P., Ofek, E., Meyer, C., Baatarsukh, K., Strejek, H., Qian, J., Freedman, J., Figueira, R., Sokolik, M., Bachem, O., Lin, R., Kharrat, D., Hidey, C., Xu, P., Duan, D., Li, Y., Ersoy, M., Everett, R., Cen, K., Santamaria-Fernandez, R., Taubenfeld, A., Mackinnon, I., Deng, L., Zablotskaia, P., Viswanadha, S., Goel, S., Yates, D., Deng, Y., Choy, P., Chen, M., Sinha, A., Mossin, A., Wang, Y., Szlam, A., Hao, S., Rubenstein, P. K., Toksoz-Exley, M., Aperghis, M., Zhong, Y., Ahn, J., Isard, M., Lacombe, O., Luisier, F., Anastasiou, C., Kalley, Y., Prabhu, U., Dunleavy, E., Bijwadia, S., Mao-Jones, J., Chen, K., Pasumarthi, R., Wood, E., Dostmohamed, A., Hurley, N., Simsa, J., Parrish, A., Pajarskas, M., Harvey, M., Skopek, O., Kochinski, Y., Rey, J., Rieser, V., Zhou, D., Lee, S. J., Acharya, T., Li, G., Jiang, J., Zhang, X., Gipson, B., Mahintorabi, E., Gelmi, M., Khajehnouri, N., Yeh, A., Lee, K., Matthey, L., Baker, L., Pham, T., Fu, H., Pak, A., Gupta, P., Vasconcelos, C., Sadovsky, A., Walker, B., Hsiao, S., Zochbauer, P., Marzoca, A., Velan, N., Zeng, J., Baechler, G., Driess, D., Jain, D., Huang, Y., Tao, L., Maggs, J., Levine, N., Schneider, J., Gemzer, E., Petit, S., Han, S., Fisher, Z., Zelle, D., Biles, C., Ie, E., Fadeeva, A., Liu, C., Franco, J. V., Collister, A., Zhang, H., Wang, R., Zhao, R., Kieliger, L., Shuster, K., Zhu, R., Gong, B., Chan, L., Sun, R., Basu, S., Zimmermann, R., Hayes, J., Bapna, A., Snoek, J., Yang, W., Datta, P., Abdallah, J. A., Kilgour, K., Li, L., Mah, S. Q., Jun, Y., Rivire, M., Karmarkar, A., Spalink, T., Huang, T., Gonzalez, L., Tran, D.-H., Nowak, A., Palowitch, J., Chadwick, M., Talius, E., Mehta, H., Sellam, T., Frnken, P., Nicosia, M., He, K., Kini, A., Amos, D., Basu, S., Jobe, H., Shaw, E., Xu, Q., Evans, C., Ikeda, D., Yan, C., Jin, L., Wang, L., Yadav, S., Labzovsky, I., Sampath, R., Ma, A., Schumann, C., Siddhant, A., Shah, R., Youssef, J., Agarwal, R., Dabney, N., Tonioni, A., Ambar, M., Li, J., Guyon, I., Li, B., Soergel, D., Fang, B., Karadzhov, G., Udrescu, C., Trinh, T., Raunak, V., Noury, S., Guo, D., Gupta, S., Finkelstein, M., Petek, D., Liang, L., Billock, G., Sun, P., Wood, D., Song, Y., Yu, X., Matejovicova, T., Cohen, R., Andra, K., D'Ambrosio, D., Deng, Z., Nallatamby, V., Songhori, E., Dangovski, R., Lampinen, A., Botadra, P., Hillier, A., Cao, J., Baddi, N., Kuncoro, A., Yoshino, T., Bhagatwala, A., Ranzato, M., Schaeffer, R., Liu, T., Ye, S., Sarvana, O., Nham, J., Kuang, C., Gao, I., Baek, J., Mittal, S., Wahid, A., Gergely, A., Ni, B., Feldman, J., Muir, C., Lamblin, P., Macherey, W., Dyer, E., Kilpatrick, L., Campos, V., Bhutani, M., Fort, S., Ahmad, Y., Severyn, A., Chatziprimou, K., Ferludin, O., Dimarco, M., Kusupati, A., Heyward, J., Bahir, D., Villela, K., Millican, K., Marcus, D., Bahargam, S., Unlu, C., Roth, N., Wei, Z., Gopal, S., Ghoshal, D., Lee, E., Lin, S., Lees, J., Lee, D., Hosseini, A., Fan, C., Neel, S., Wu, M., Altun, Y., Cai, H., Piqueras, E., Woodward, J., Bissacco, A., Haykal, S., Bordbar, M., Sundaram, P., Hodkinson, S., Toyama, D., Polovets, G., Myers, A., Sinha, A., Levinboim, T., Krishnakumar, K., Chhaparia, R., Sholokhova, T., Gundavarapu, N. B., Jawahar, G., Qureshi, H., Hu, J., Momchev, N., Rahtz, M., Wu, R., S, A. P., Dhamdhere, K., Guo, M., Gupta, U., Eslami, A., Schain, M., Blokzijl, M., Welling, D., Orr, D., Bolelli, L., Perez-Nieves, N., Sirotenko, M., Prasad, A., Kar, A., Pigem, B. D. B., Terzi, T., Weisz, G., Ghosh, D., Mavalankar, A., Madeka, D., Daugaard, K., Adam, H., Shah, V., Berman, D., Tran, M., Baker, S., Andrejczuk, E., Chole, G., Raboshchuk, G., Mirzazadeh, M., Kagohara, T., Wu, S., Schallhart,

C., Orlando, B., Wang, C., Rrustemi, A., Xiong, H., Liu, H., Vezer, A., Ramsden, N., Chang, S.-y., Mudgal, S., Li, Y., Vieillard, N., Hoshen, Y., Ahmad, F., Slone, A., Hua, A., Potikha, N., Rossini, M., Stritar, J., Prakash, S., Wang, Z., Dong, X., Nazari, A., Nehoran, E., Tekelioglu, K., Li, Y., Badola, K., Funkhouser, T., Li, Y., Yerram, V., Ganeshan, R., Formoso, D., Langner, K., Shi, T., Li, H., Yamamori, Y., Panda, A., Saade, A., Scarpati, A. S., Breaux, C., Carey, C. J., Zhou, Z., Hsieh, C.-J., Bridgers, S., Butryna, A., Gupta, N., Tulsyan, V., Woo, S., Eltyshev, E., Grathwohl, W., Parks, C., Benjamin, S., Panigrahy, R., Dodhia, S., Freitas, D. D., Sauer, C., Song, W., Alet, F., Tolins, J., Paduraru, C., Zhou, X., Albert, B., Zhang, Z., Shu, L., Bansal, M., Nguyen, S., Globerson, A., Xiao, O., Manyika, J., Hennigan, T., Rong, R., Matak, J., Bakalov, A., Sharma, A., Sinopalnikov, D., Pierson, A., Roller, S., Brown, G., Gao, M., Fukuzawa, T., Ghafouri, A., Vassigh, K., Barr, I., Wang, Z., Korsun, A., Jayaram, R., Ren, L., Zaman, T., Khan, S., Lunts, Y., Deutsch, D., Uthus, D., Katz, N., Samsikova, M., Khalifa, A., Sethi, N., Sun, J., Tang, L., Alon, U., Luo, X., Yu, D., Nayyar, A., Petrini, B., Truong, W., Hellendoorn, V., Chinaev, N., Alberti, C., Wang, W., Hu, J., Mirrokni, V., Balashankar, A., Aharon, A., Mehta, A., Iscen, A., Kready, J., Manning, L., Mohananey, A., Chen, Y., Tripathi, A., Wu, A., Petrovski, I., Hwang, D., Baeuml, M., Chandrakaladharan, S., Liu, Y., Coaguila, R., Chen, M., Ma, S., Tafti, P., Tatineni, S., Spitz, T., Ye, J., Vicol, P., Rosca, M., Puigdomnech, A., Yahav, Z., Ghemawat, S., Lin, H., Kirk, P., Nabulsi, Z., Brin, S., Bohnet, B., Caluwaerts, K., Veerubhotla, A. S., Zheng, D., Dai, Z., Petrov, P., Xu, Y., Mehran, R., Xu, Z., Zintgraf, L., Choi, J., Hombaiah, S. A., Thoppilan, R., Reddi, S., Lew, L., Li, L., Webster, K., Sawhney, K. P., Lamprou, L., Shakeri, S., Lunayach, M., Chen, J., Bagri, S., Salcianu, A., Chen, Y., Donchev, Y., Magister, C., Nrly, S., Rodrigues, V., Izo, T., Noga, H., Zou, J., Kppe, T., Zhou, W., Lee, K., Long, X., Eisenbud, D., Chen, A., Schenck, C., To, C. M., Zhong, P., Taropa, E., Truong, M., Levy, O., Martins, D., Zhang, Z., Semturs, C., Zhang, K., Yakubovich, A., Moreno, P., McConnaughey, L., Lu, D., Redmond, S., Weerts, L., Bitton, Y., Refice, T., Lacasse, N., Conmy, A., Tallec, C., Odell, J., Forbes-Pollard, H., Socala, A., Hoech, J., Kohli, P., Walton, A., Wang, R., Sazanovich, M., Zhu, K., Kapishnikov, A., Galt, R., Denton, M., Murdoch, B., Sikora, C., Mohamed, K., Wei, W., First, U., McConnell, T., Cobo, L. C., Qin, J., Avrahami, T., Balle, D., Watanabe, Y., Louis, A., Kraft, A., Ariafar, S., Gu, Y., Rives, E., Yoon, C., Rusu, A., Cobon-Kerr, J., Hahn, C., Luo, J., Yuvein, Zhu, Ahuja, N., Benenson, R., Kaufman, R. L., Yu, H., Hightower, L., Zhang, J., Ni, D., Hendricks, L. A., Wang, G., Yona, G., Jain, L., Barrio, P., Bhupatiraju, S., Velusamy, S., Dafoe, A., Riedel, S., Thomas, T., Yuan, Z., Bellaiche, M., Panthaplackel, S., Kloboves, K., Jauhari, S., Akbulut, C., Davchev, T., Gladchenko, E., Madras, D., Chuklin, A., Hill, T., Yuan, Q., Madhavan, M., Leonhard, L., Scandinaro, D., Chen, Q., Niu, N., Douillard, A., Damoc, B., Onoe, Y., Pedregosa, F., Bertsch, F., Leichner, C., Pagadora, J., Malmaud, J., Ponda, S., Twigg, A., Duzhyi, O., Shen, J., Wang, M., Garg, R., Chen, J., Evci, U., Lee, J., Liu, L., Kojima, K., Yamaguchi, M., Rajendran, A., Piergiovanni, A. J., Rajendran, V. K., Fornoni, M., Ibagon, G., Ragan, H., Khan, S. M., Blitzer, J., Bunner, A., Sun, G., Kosakai, T., Lundberg, S., Elue, N., Guu, K., Park, S. K., Park, J., Narayanaswamy, A., Wu, C., Mudigonda, J., Cohn, T., Mu, H., Kumar, R., Graesser, L., Zhang, Y., Killam, R., Zhuang, V., Gimnez, M., Jishi, W. A., Ley-Wild, R., Zhai, A., Osawa, K., Cedillo, D., Liu, J., Upadhyay, M., Sieniek, M., Sharma, R., Paine, T., Angelova, A., Addepalli, S., Parada, C., Majumder, K., Lamp, A., Kumar, S., Deng, X., Myaskovsky, A., Saboli, T., Dudek, J., York, S., Quitry, F. d. C., Nie, J., Cattle, D., Gunjan, A., Piot, B., Khawaja, W., Bang, S., Wang, S., Khodadadeh, S., R, R., Rawlani, P., Powell, R., Lee, K., Griesser, J., Oh, G. S., Magalhaes, C., Li, Y., Tokumine, S., Vogel, H. N., Hsu, D., BC, A., Jindal, D., Cohen, M., Yang, Z., Yuan, J., Cesare, D. d., Bruguier, T., Xu, J., Roy, M., Jacovi, A., Belov, D., Arya, R., Meadowlark, P., Cohen-Ganor, S., Ye, W., Morris-Suzuki, P., Banzal, P., Song, G., Ponnuramu, P., Zhang, F., Scrivener, G., Zaiem, S., Rochman, A. R., Han, K., Ghazi, B., Lee, K., Drath, S., Suo, D., Girgis, A., Shenoy, P., Nguyen, D., Eck, D., Gupta, S., Yan, L., Carreira, J., Gulati, A., Sang, R., Mirylenka, D., Cooney, E., Chou, E., Ling, M., Fan, C., Coleman, B., Tubone, G., Kumar, R., Baldridge, J., Hernandez-Campos, F., Lazaridou, A., Besley, J., Yona, I., Bulut, N., Wellens, Q., Pierigiovanni, A. J., George, J., Green, R., Han, P., Tao, C., Clark, G., You, C., Abdolmaleki, A., Fu, J., Chen, T., Chaugule, A., Chandorkar, A., Rahman, A., Thompson, W., Koanantakool, P., Bernico, M., Ren, J., Vlasov, A., Vassilvitskii, S., Kula, M., Liang, Y., Kim, D., Huang, Y., Ye, C., Lepikhin, D., and Helmholz, W. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities, October 2025. URL http://arxiv.org/abs/2507.06261. arXiv:2507.06261 [cs].

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Ding, H., Xin, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Wang, J., Chen, J., Yuan, J., Qiu, J., Li, J., Cai, J. L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L.,

Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R. J., Jin, R. L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S. S., Zhou, S., Wu, S., Ye, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Zhao, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W. L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X. Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y. K., Wang, Y. Q., Wei, Y. X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y. X., Xu, Y., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., and Zhang, Z. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, January 2025. URL /. arXiv:2501.12948 [cs].

Dekoninck, J., Petrov, I., Minchev, K., Balunovic, M., Vechev, M., Marinov, M., Drencheva, M., Konova, L., Shumanov, M., Tsvetkov, K., Drenchev, N., Todorov, L., Nikolova, K., Georgiev, N., Kalinkova, V., and Ismoldayev, M. The Open Proof Corpus: A Large-Scale Study of LLM-Generated Mathematical Proofs, June 2025. URL http://arxiv.org/abs/2506.21621. arXiv:2506.21621 [cs].

Haldar, R. and Hockenmaier, J. Rating Roulette: Self-Inconsistency in LLM-As-A-Judge Frameworks, October 2025. URL http://arxiv.org/abs/2510.27106. arXiv:2510.27106 [cs].

Huang, Y. and Yang, L. F. Gemini 2.5 Pro Capable of Winning Gold at IMO 2025, July 2025. URL http://arxiv.org/abs/2507.15855. arXiv:2507.15855 [cs].

Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., Hasin, M., Jawhar, S., Kinniment, M., Rush, N., Arx, S. V., Bloom, R., Broadley, T., Du, H., Goodrich, B., Jurkovic, N., Miles, L. H., Nix, S., Lin, T., Parikh, N., Rein, D., Sato, L. J. K., Wijk, H., Ziegler, D. M., Barnes, E., and Chan, L. Measuring AI Ability to Complete Long Tasks, March 2025. URL http://arxiv.org/abs/2503.14499. arXiv:2503.14499 [cs].

Li, Q., Dou, S., Shao, K., Chen, C., and Hu, H. Evaluating Scoring Bias in LLM-as-a-Judge, August 2025. URL http://arxiv.org/abs/2506.22316. arXiv:2506.22316 [cs].

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment, May 2023. URL http://arxiv.org/abs/2303.16634. arXiv:2303.16634 [cs].

Liu, Y., Huang, Y., Wang, Y., Li, P., and Liu, Y. AI Mathematician: Towards Fully Automated Frontier Mathematical Research, May 2025. URL http://arxiv.org/abs/2505.22451. arXiv:2505.22451 [cs].

Luong, T., Hwang, D., Nguyen, H. H., Ghiasi, G., Chervonyi, Y., Seo, I., Kim, J., Bingham, G., Lee, J., Mishra, S., Zhai, A., Hu, C. H., Michalewski, H., Kim, J., Ahn, J., Bae, J., Song, X., Trinh, T. H., Le, Q. V., and Jung, J. Towards Robust Mathematical Reasoning, November 2025. URL http://arxiv.org/abs/2511.01846. arXiv:2511.01846 [cs].

Mahdavi, H., Mahdavinia, P., Malek, S., Mohammadipour, P., Hashemi, A., Daliri, M., Farhadi, A., Khasahmadi, A., Mireshghallah, N., and Honavar, V. RefGrader: Automated Grading of Mathematical Competition Proofs using Agentic Workflows, October 2025a. URL http://arxiv.org/abs/2510.09021. arXiv:2510.09021 [cs].

Mahdavi, S., Kisacanin, B., Toshniwal, S., Du, W., Moshkov, I., Armstrong, G., Liao, R., Thrampoulidis, C., and Gitman, I. Scaling Generative Verifiers For Natural Language Mathematical Proof Verification And Selection, November 2025b. URL http://arxiv.org/abs/2511.13027. arXiv:2511.13027 [cs].

OpenAI, Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., Iftimie, A., Karpenko, A., Passos, A. T., Neitz, A., Prokofiev, A., Wei, A., Tam, A., Bennett, A., Kumar, A., Saraiva, A., Vallone, A., Duberstein, A., Kondrich, A., Mishchenko, A., Applebaum, A., Jiang, A., Nair, A., Zoph, B., Ghorbani, B., Rossen, B., Sokolowsky, B., Barak, B., McGrew, B., Minaiev, B., Hao, B., Baker, B., Houghton, B., McKinzie, B., Eastman, B., Lugaresi, C., Bassin, C., Hudson, C., Li, C. M., Bourcy, C. d., Voss, C., Shen, C., Zhang, C., Koch, C., Orsinger, C., Hesse, C., Fischer, C., Chan, C., Roberts, D., Kappler, D., Levy, D., Selsam, D., Dohan, D., Farhi, D., Mely, D., Robinson, D., Tsipras, D., Li, D., Oprica, D., Freeman, E., Zhang, E., Wong, E., Proehl, E., Cheung, E., Mitchell, E., Wallace, E., Ritter, E., Mays, E., Wang, F., Such, F. P., Raso, F., Leoni, F., Tsimpourlas, F., Song, F., Lohmann, F. v., Sulit, F., Salmon, G., Parascandolo, G., Chabot, G., Zhao, G., Brockman, G., Leclerc, G., Salman, H., Bao, H., Sheng, H., Andrin, H., Bagherinezhad, H., Ren, H.,

Lightman, H., Chung, H. W., Kivlichan, I., O'Connell, I., Osband, I., Gilaberte, I. C., Akkaya, I., Kostrikov, I., Sutskever, I., Kofman, I., Pachocki, J., Lennon, J., Wei, J., Harb, J., Twore, J., Feng, J., Yu, J., Weng, J., Tang, J., Yu, J., Candela, J. Q., Palermo, J., Parish, J., Heidecke, J., Hallman, J., Rizzo, J., Gordon, J., Uesato, J., Ward, J., Huizinga, J., Wang, J., Chen, K., Xiao, K., Singhal, K., Nguyen, K., Cobbe, K., Shi, K., Wood, K., Rimbach, K., Gu-Lemberg, K., Liu, K., Lu, K., Stone, K., Yu, K., Ahmad, L., Yang, L., Liu, L., Maksin, L., Ho, L., Fedus, L., Weng, L., Li, L., McCallum, L., Held, L., Kuhn, L., Kondraciuk, L., Kaiser, L., Metz, L., Boyd, M., Trebacz, M., Joglekar, M., Chen, M., Tintor, M., Meyer, M., Jones, M., Kaufer, M., Schwarzer, M., Shah, M., Yatbaz, M., Guan, M. Y., Xu, M., Yan, M., Glaese, M., Chen, M., Lampe, M., Malek, M., Wang, M., Fradin, M., McClay, M., Pavlov, M., Wang, M., Wang, M., Murati, M., Bavarian, M., Rohaninejad, M., McAleese, N., Chowdhury, N., Chowdhury, N., Ryder, N., Tezak, N., Brown, N., Nachum, O., Boiko, O., Murk, O., Watkins, O., Chao, P., Ashbourne, P., Izmailov, P., Zhokhov, P., Dias, R., Arora, R., Lin, R., Lopes, R. G., Gaon, R., Miyara, R., Leike, R., Hwang, R., Garg, R., Brown, R., James, R., Shu, R., Cheu, R., Greene, R., Jain, S., Altman, S., Toizer, S., Toyer, S., Miserendino, S., Agarwal, S., Hernandez, S., Baker, S., McKinney, S., Yan, S., Zhao, S., Hu, S., Santurkar, S., Chaudhuri, S. R., Zhang, S., Fu, S., Papay, S., Lin, S., Balaji, S., Sanjeev, S., Sidor, S., Broda, T., Clark, A., Wang, T., Gordon, T., Sanders, T., Patwardhan, T., Sottiaux, T., Degry, T., Dimson, T., Zheng, T., Garipov, T., Stasi, T., Bansal, T., Creech, T., Peterson, T., Eloundou, T., Qi, V., Kosaraju, V., Monaco, V., Pong, V., Fomenko, V., Zheng, W., Zhou, W., McCabe, W., Zaremba, W., Dubois, Y., Lu, Y., Chen, Y., Cha, Y., Bai, Y., He, Y., Zhang, Y., Wang, Y., Shao, Z., and Li, Z. OpenAI o1 System Card, December 2024. URL http://arxiv.org/abs/2412.16720. arXiv:2412.16720 [cs].

Pandit, S., Xu, A., Nguyen, X.-P., Ming, Y., Xiong, C., and Joty, S. Hard2Verify: A Step-Level Verification Benchmark for Open-Ended Frontier Math, October 2025. URL http://arxiv.org/abs/2510.13744. arXiv:2510.13744 [cs].

Shi, W. and Jin, X. Heimdall: test-time scaling on the generative verification, April 2025. URL http://arxiv.org/abs/2504.10337. arXiv:2504.10337 [cs].

Sinha, A., Arun, A., Goel, S., Staab, S., and Geiping, J. The Illusion of Diminishing Returns: Measuring Long Horizon Execution in LLMs, September 2025. URL http://arxiv.org/abs/2509.09677. arXiv:2509.09677 [cs].

Team, K., Bai, Y., Bao, Y., Chen, G., Chen, J., Chen, N., Chen, R., Chen, Y., Chen, Y., Chen, Y., Chen, Z., Cui, J., Ding, H., Dong, M., Du, A., Du, C., Du, D., Du, Y., Fan, Y., Feng, Y., Fu, K., Gao, B., Gao, H., Gao, P., Gao, T., Gu, X., Guan, L., Guo, H., Guo, J., Hu, H., Hao, X., He, T., He, W., He, W., Hong, C., Hu, Y., Hu, Z., Huang, W., Huang, Z., Huang, Z., Jiang, T., Jiang, Z., Jin, X., Kang, Y., Lai, G., Li, C., Li, F., Li, H., Li, M., Li, W., Li, Y., Li, Y., Li, Z., Li, Z., Lin, H., Lin, X., Lin, Z., Liu, C., Liu, C., Liu, H., Liu, J., Liu, J., Liu, L., Liu, S., Liu, T. Y., Liu, T., Liu, W., Liu, Y., Liu, Y., Liu, Y., Liu, Y., Liu, Z., Lu, E., Lu, L., Ma, S., Ma, X., Ma, Y., Mao, S., Mei, J., Men, X., Miao, Y., Pan, S., Peng, Y., Qin, R., Qu, B., Shang, Z., Shi, L., Shi, S., Song, F., Su, J., Su, Z., Sun, X., Sung, F., Tang, H., Tao, J., Teng, Q., Wang, C., Wang, D., Wang, F., Wang, H., Wang, J., Wang, J., Wang, J., Wang, S., Wang, S., Wang, Y., Wang, Y., Wang, Y., Wang, Y., Wang, Y., Wang, Z., Wang, Z., Wang, Z., Wei, C., Wei, Q., Wu, W., Wu, X., Wu, Y., Xiao, C., Xie, X., Xiong, W., Xu, B., Xu, J., Xu, J., Xu, L. H., Xu, L., Xu, S., Xu, W., Xu, X., Xu, Y., Xu, Z., Yan, J., Yan, Y., Yang, X., Yang, Y., Yang, Z., Yang, Z., Yang, Z., Yao, H., Yao, X., Ye, W., Ye, Z., Yin, B., Yu, L., Yuan, E., Yuan, H., Yuan, M., Zhan, H., Zhang, D., Zhang, H., Zhang, W., Zhang, X., Zhang, Y., Zhang, Y., Zhang, Y., Zhang, Y., Zhang, Y., Zhang, Y., Zhang, Z., Zhao, H., Zhao, Y., Zheng, H., Zheng, S., Zhou, J., Zhou, X., Zhou, Z., Zhu, Z., Zhuang, W., and Zu, X. Kimi K2: Open Agentic Intelligence, July 2025. URL http://arxiv.org/abs/2507.20534. arXiv:2507.20534 [cs].

Varambally, S., Voice, T., Sun, Y., Chen, Z., Yu, R., and Ye, K. Hilbert: Recursively Building Formal Proofs with Informal Reasoning, September 2025. URL http://arxiv.org/abs/2509.22819. arXiv:2509.22819 [cs].

Xu, Y., Chakraborty, T., Sharma, S., Nunes, L., Kcman, E., Lu, S., and Chandra, R. Direct Reasoning Optimization: LLMs Can Reward And Refine Their Own Reasoning for Open-Ended Tasks, June 2025. URL http://arxiv.org/abs/2506.13351. arXiv:2506.13351 [cs].

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. Qwen3 Technical Report, May 2025. URL http://arxiv.org/abs/2505.09388. arXiv:2505.09388 [cs].

Yu, T., Ji, B., Wang, S., Yao, S., Wang, Z., Cui, G., Yuan,

L., Ding, N., Yao, Y., Liu, Z., Sun, M., and Chua, T.-S. RLPR: Extrapolating RLVR to General Domains without Verifiers, June 2025. URL http://arxiv.org/abs/2506.18254. arXiv:2506.18254 [cs].

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, December 2023. URL http://arxiv.org/abs/2306.05685. arXiv:2306.05685 [cs].

Zuo, Y., Zhang, K., Qu, S., Sheng, L., Zhu, X., Qi, B., Sun, Y., Cui, G., Ding, N., and Zhou, B. TTRL: Test-Time Reinforcement Learning, April 2025. URL http://arxiv.org/abs/2504.16084. arXiv:2504.16084 [cs].