Competition and Judging Handbook

Open Problems Hack at MIT

September 2026

Introduction This is the official handbook of the Open Problems Hack at MIT. The scoring, judging and competitors will be regulated according to this document and its rules will be applied as best as the esteemed jury can guarantee them.

Participant links: Event website and updates | Registration and host contact. The registration page's *Contact the Host* option is available for questions.

1 Objective and competition structure

The objective is to produce the largest possible body of durable mathematics within one concentrated week: solutions to genuine open problems, meaningful formalized partial advances, and useful formalizations of known mathematics. The competition tests human research, AI models, formal systems, and computational harnesses at research scale.

Only formalized results count. Every accepted claim must pass a pinned, approved formal checker and mathematical review of its exact statement, novelty, scope, and attribution. Informal proofs, numerical patterns, and model confidence do not score, which is a necessity due to the volume of the competition and the goal of automating mathematics research.

A checker validates a formal statement under its declared assumptions; reviewers must still establish that it faithfully represents the claimed mathematics. There is no mandatory oral defense.

Participants may work worldwide, with any AI provider, other tools, or no AI, and submit as many distinct results as they can finish. There are three modalities, with two branches in the third:

Mode	Eligible contribution	Treatment
M1	Solutions or formalized advances on the frozen curated focus set.	Original-problem score with a focus bonus of 1.1x.
M2	New, faithful, useful, reusable formalizations of known mathematics.	Separate leaderboard by number of distinct accepted formalization families. Not counted as open problems solved.
M3A	Solutions or formalized advances on other eligible original open problems, including newly admitted targets.	Original-problem score without the focus bonus.
M3B	Solutions or formalized advances on nontrivial, parent-linked variations of Open Mathematics problems.	Own difficulty and a 1/2 modality multiplier.

Each contribution has one primary modality; the same work cannot earn duplicate modality credit. Proposing a problem alone earns no core score. Any separate curation or formulation distinction must be announced in advance.

2 Calendar, access, and entrant classes

2.1 Schedule and eligibility window

- **Sunday, 27 September, noon Eastern / 9 a.m. Pacific:** hybrid opening at MIT CSAIL, with remote access, onboarding, team formation, tool/submission demonstrations, and initial work.
  The talks will follow the noon opening to give contestants context and motivate their work.

- **Monday, 28 September–Friday, 2 October:** research, rolling problem admission, submissions, and review. Daily checkpoints and office hours support participants but do not score.
- **Friday, 2 October:** all score-bearing mathematics and formal artifacts must be complete and submitted before 00:00 EDT on Saturday, 3 October (9 p.m. PDT on Friday, 2 October). Primary judging takes place on Friday and the immediately following days.

Anyone may participate remotely or join after Sunday, subject to the same Friday requirements. Neither MIT attendance nor attendance at the opening is required. The common mathematical baseline is the opening/status-freeze time, noon Eastern on 27 September, even for late entrants.

After Friday, organizers may complete reviews of already-finished artifacts, deterministic reruns, deduplication, appeals, attribution, and publication assembly. A subsequent results ceremony is provisionally expected around one to two weeks later, with details announced when ready. This is not a second work period, a second award horizon, or a multi-month verification round.

2.2 One entrant, one class

An entrant comprises **one to four named humans**. Company status takes precedence over headcount:

- **Individual:** one person entering independently of a company.
- **Team:** two to four people entering independently of a company.
- **Company:** one to four people officially representing a company, doing company-directed or paid company work, or using material company-funded compute or other dedicated company resources, whether public or proprietary.

Academic affiliation and student status are descriptive tags, not additional ranking classes. Ordinary employment alone does not make a personal-time, public-tool entry a company entry. Public tools and organizer-allocated sponsor credits do not by themselves trigger company classification. Disclose material private compute, models, infrastructure, data, or funding; the chair records borderline classifications before scoring. Mixed entries with material company support belong in the company class. These classes do not guarantee equal resources within a class.

Each person belongs to one active score-bearing entrant. Record roster and class at the first eligible submission. Chair-approved corrections or mergers must preserve attribution, stay within four members, transfer prior credit exclusively to the resulting entry, and leave an audit trail; no result or person is ranked twice. Company rankings compare registered company entries, not the combined output of all employees. Separate entries must remain genuinely independent.

OpenAI or other provider credits, if available, are allocated case by case. No sponsorship, credit allocation, or provider access is guaranteed or required for eligibility.

3 Problem infrastructure, admission, and difficulty

3.1 Public corpus and canonical records

The OPDP begins from a versioned public corpus, expanding the previously classified set with MathDB. Open Mathematics supplies the broader problem infrastructure. Import, deduplication, provenance checks, and status review are ongoing; database inclusion does not certify that a problem remains open. Publish source snapshots, import coverage, unresolved duplicates, ambiguities, and the frozen focus list rather than an unsupported final corpus count.

Every scoreable target requires a canonical record: problem/family IDs; exact statement and success criterion; source/snapshot; open-status evidence; modality; parent links; domain/steward; frozen difficulty and source scale; focus status; exclusions; registration timestamp; and version history. Material changes

to hypotheses, quantifiers, or conclusions require a new version and reassessment. Clerical corrections are logged; the exact submitted target remains identifiable.

New problems may be admitted throughout the week. Admission requires a precise mathematical claim, assessable difficulty, a suitable reviewer, and no redundant family credit. M1/M3A targets must predate the event and remain open at the status freeze; new event-created targets belong in M3B, not M3A. Known targets with new formalizations belong in M2. M3B variations must be nontrivial and not already solved at the freeze. Without qualified review, admission is provisional; no final score is awarded until status, family, and difficulty are confirmed.

Openness is judged against public mathematical knowledge at the common status freeze, including for targets admitted later. A previously published solution discovered after admission requires reclassification, not open-problem credit merely because participants were unaware of it.

3.2 Known mathematics and variations

For M2, identify the source theorem, prior formal libraries searched, and the genuinely new formal contribution. Renaming declarations, re-exporting library results, or restating an existing formalization is not a new accomplishment.

For M3B, supply the canonical parent, exact variation, why it is nontrivial, and prior knowledge of an answer or method. Copies, immediate corollaries, and cosmetic parameter changes do not create independent scoreable problems. A variation receives its own difficulty, not its parent's. If review establishes equivalence to the original problem, the chair reclassifies it before scoring. Attribute formulation and solution separately where appropriate.

3.3 One frozen competition-difficulty scale

Use the published OPDP difficulty score supplied to contestants on the 0–1000 scale directly as $D(P)$, without rescaling or substituting another Atlas dimension. This is the *competition difficulty* from the OPDP Atlas developed by Alejandro Zarzuelo Urdiales at `https://github.com/alejandrozu/ulam-opdp-difficulty-atlas`. Problems that are not yet incorporated into the atlas will be evaluated with the same methodology and added automatically to its corpus. The problems have to be presented with a fully formalized statement and sources.

New targets receive three independent assessments under the same methodology; use the median. A spread exceeding 150 points requires discussion and documented consensus or majority resolution. Where feasible, assess a new statement before reading its solution to limit hindsight bias.

4 Scoring and ranked views

4.1 Quadratic difficulty; additive output

For entrant T and an accepted score-eligible family P, set

$$\boxed{F_{T,P} = b_{T,P}\, m(P)\, p_{T,P}\, \frac{D(P)^2}{1000}} \qquad \left(\gamma = 2, \quad V(D) = 1000 \left(\frac{D}{1000}\right)^2\right).$$

The base multiplier is $b_{T,P} = 1.1$ for the frozen M1 focus set and $b_{T,P} = 1$ otherwise. The jury may award an additional bonus for unusually consequential results; the entrant may make a short case of mathematical importance, particularly connections to other mathematics and potential new insights. Any such bonus is included in the result's recorded $b_{T,P}$, with its amount and rationale published in the accepted review, and affects both rankings below.

The *modality multiplier* is $m = 1$ for M1/M3A and $m = 1/2$ for M3B. The reviewed progress coefficient $p \in [0, 1]$ applies to the entrant's accepted union of work within that family. Only difficulty is squared: progress and multipliers remain linear.

Publish both rankings separately for **individuals, teams, and company entries**:

$$\boxed{S_T = \sum_{P \in \mathcal{F}_T} F_{T,P}} \quad \text{Total accepted output,}$$

$$\boxed{A_T = \max_{P \in \mathcal{F}_T} F_{T,P}} \quad \text{Biggest accomplishment.}$$

Thus there are six open-problem ranked views, plus the separate M2 rankings below. $\mathcal{F}_T$ contains distinct accepted M1/M3A/M3B families; $A_T = 0$ when it is empty. Biggest accomplishment means the *highest-scoring accepted family contribution*, not the largest bundle of unrelated results or a jury's universal judgment of significance. Show the winning family, exact accepted claim, modality, and completion status next to A_T.

Revisions and overlapping subclaims **replace the previous family score**; they do not stack. Distinct families add without a submission cap, difficulty-band cap, or harmonic discount. Store inputs as exact decimals at published precision and calculate scores exactly; round only for display, never to determine ranks or ties. The same final recorded inputs generate both views.

M2 has a separate leaderboard in each entrant class, ranked by the number of distinct accepted formalization families (higher is better). A family is one registered source theorem or admitted reusable formalization target; revisions, overlapping declarations, and supporting lemmas for that target count together, not separately. Each counted contribution must be new, faithful, useful, and reusable. The independence and deduplication rules below apply. M2 counts do not enter S_T, A_T, their mathematical tie-break criteria, or open-problems-solved counts.

4.2 Examples and ties

For complete, non-focus original results ($b = m = p = 1$):

Difficulty D	400	600	800	900	1000
Base points $D^2/1000$	160	360	640	810	1000

Two distinct complete $D = 600$ results give $S = 720$ and $A = 360$; one complete $D = 800$ result gives $S = A = 640$. The first entrant leads on volume, the second on biggest accomplishment. A $D = 800$ result accepted at $p = 0.5$ earns 320 points; an M3B result with the same D, p earns 160. Improving that original family from $p = 0.5$ to $p = 1$ ends at 640, not 960, before any focus bonus.

For tied totals, compare in order: score from complete original M1/M3A solutions; number of those solutions; then score from original formalized partial advances. **Time is the final tie-breaker in every leaderboard**: use the earlier server-recorded submission timestamp, to millisecond precision, of the eligible version that last increased the entrant's final value in that view (S_T, A_T, or the M2 count). Review times, rejected versions, and non-score-changing uploads do not count. Total score does not break a biggest-accomplishment tie. Exact timestamp ties at the recorded precision share rank. Apart from the explicit additional bonus above, jury distinctions for significance, elegance, verification, or formulation remain separate and do not change numerical standings.

5 Formalized partial credit and special cases

Progress measures accepted logical achievement, not effort, presentation, compute spent, or confidence that the remainder will work. Judges assess coverage, generality, pivotality, novelty, and the exact residual obstacle. A coefficient is a scoring judgment, not a literal percentage of a theorem.

Band	p	**Accepted contribution**
P0	0	No eligible formalized advance.
P1	.01–.05	Narrow certified computation, obstruction, negative theorem, or reusable fact.
P2	.05–.15	Nontrivial lemma, reduction, reformulation, finite classification, or improvement short of the central obstacle.
P3	.15–.35	Meaningful infinite family, special case, bound, or structural advance removing a recognized obstacle.
P4	.35–.60	Major cases or components carrying a substantial part of the logical burden.
P5	.60–.85	Most of the target resolved; the remainder is explicit, localized, and materially smaller.
P6	.85–.95	Near-complete accepted result with a genuine nontrivial remainder.
Full	1	The exact registered target is completely proved or validly refuted.

No coefficient strictly between .95 and 1 is allowed. Start at the lower end of the appropriate band unless a written rationale supports more. Two independent judges propose a band and coefficient. If they agree on the band and differ by at most .10, they discuss a joint decision; a larger difference or band disagreement requires a third reviewer and recorded chair resolution. Overlapping boundary values require an explicit band choice, not two awards.

A claimed complete solution that fails acceptance earns only credit for separately specified, formally accepted subclaims, if any; it never receives $p = 1$ on expectation of repair.

- **Counterexamples:** full credit requires a formal example satisfying the registered hypotheses and refuting the actual universal claim. Refuting an auxiliary assertion is scored only for that narrower contribution.
- **Computation:** exhaustive finite results require a formal certificate or verified procedure and may justify credit at the appropriate scope, often P1–P3. Sampling does not prove an uncovered infinite statement.
- **Negative results:** failure of a particular method usually belongs in P1 or low P2; broad formal impossibility theorems may warrant more according to their scope.
- **Prior mathematics:** known results receive no new open-problem credit. New formalizations go to M2; assess genuinely new remainders after removing prior overlap.

6 Families, independence, collaboration, and baseline

Score mathematics, not packaging. Equivalent formulations, trivial variants, fragments, and overlapping lemmas belong to one family where appropriate. A theorem settling several targets normally earns one highest eligible primary score; record all implications scientifically, but award extra credit only for independently substantive additional work. Do not bundle unrelated results into a family to inflate biggest accomplishment. Family assignments, representatives, and changes require justification and logging.

Problems are nonexclusive. Participants may indicate active work on a dashboard without exposing private proofs. Materially independent, sealed solutions submitted before exposure to another entrant's work may each receive full entrant credit after independence review; the Collective Mathematical Frontier counts the result once. Timestamps, access history, and disclosure of prior exposure support that decision. After reading a solution, an alternative proof, verification, or extension earns only its

eligible distinct contribution, not automatic duplicate full credit. Joint work belongs to one registered entrant. Verification alone never duplicates the original family score; any separate recognition must be a pre-published distinction or capped separate pool.

Only identifiable mathematics contributed within the common competition window counts as event output. Pre-existing tools, libraries, and research may be used, with a timestamped baseline commit and an explicit new-work delta. Earlier mathematics remains cited prior work even if uploaded during the event. New formalization of a known theorem can qualify for M2 when the formal contribution is made during the window.

Disclose substantive external assistance and each human's contribution. The chair determines whether outside input is ordinary consultation, requires adding an author/member within the four-person limit, or makes the entry ineligible. Never omit a genuine contributor to circumvent that limit. Mergers, shared work, and authorship changes require recorded review; no individual point allocation is manufactured inside a joint entry.

7 Tools, formal verification, and Autolab Hills

7.1 Tool neutrality and reproducibility

Any AI system, proof assistant, CAS, custom agent/harness, and human-only reasoning is allowed. For each claim, disclose material model/tool names and versions, roles, material interactions, external data/libraries, private corpora, approximate compute and funding/credit sources, harness advantages, and what humans independently checked. Preserve relevant code, notebooks, seeds, and provenance. Exhaustive exploratory prompt logs are not required; do not publish credentials, personal information, or unrelated confidential material.

Replace proprietary or unavailable model conversations with a self-contained mathematical and reproducibility account sufficient for review. Model agreement, confidence, or an impressive transcript is not proof. Do not upload confidential manuscripts, review material, or another entrant's private work into unauthorized systems; organizers identify approved systems for sensitive review.

Every score-bearing artifact needs the exact formal statement, pinned prover/library versions, reproducible build instructions, declared axioms/dependencies, and generated code or certificates. No proof holes, placeholders, unavailable dependencies, or undeclared trust shortcuts are permitted. The verification chair publishes allowed trust boundaries and certificate protocols. Computer-assisted arguments need the precise reduction, code/certificate, environment, and independent rerun/check. Reviewers check correspondence with the registered claim; stronger hidden hypotheses or weaker conclusions cannot masquerade as the original solution.

We, as the jury, appeal to the common sense and good intentions of contestants to understand that the volume of submissions requires us to have a streamlined approach and that increasing the work that we have to perform, while not negatively affecting the scores, constitutes a violation of the social contract and strongly encourage teams to keep everything as self contained as possible.

7.2 Required submission route

Each entrant registers an Autolab account/workspace with one accountable owner and all human members identified, even if the platform represents a team through one owner. Use the published competition submission workflow for *every* claimed formalized result; general setup help is available in the Autolab documentation.

A *Hill* is a versioned task/evaluator; a *Climb* records an attempt sequence. Each exact score-bearing claim needs its submission ID, Hill version/tree hash, Climb link, immutable final commit, and final evaluator report. Revisions link to earlier versions. New targets must be admitted before scoring and use

the committed competition Hill where applicable. Competition-specific access, templates, namespaces, supported environments, and the outage/fallback route will be linked from the event website before submissions open; general Autolab documentation does not replace this workflow.

A passing Hill is necessary under that workflow but is not itself a mathematical proof: formal checking, statement fidelity, literature status, attribution, and review remain separate gates. The final artifact must already pass the required formal checks by Friday's cutoff; organizer reruns may follow without changing the mathematics.

8 Minimum submission packet and deadline rules

One concise packet must provide the following; linked immutable records may satisfy repeated fields:

1. **Identity/target:** submission/version, problem/family IDs, modality, roster/class, affiliations/resource classification, authors/contributions, source/snapshot, exact claim, and claimed completeness. A requested p is advisory.
2. **Artifact/proof:** Autolab owner, Hill hash/version, Climb, final commit/report, formal source, environment/reproduction commands, statement-correspondence note, derivation, definitions, references, axioms, trust dependencies, and code/certificates. The proof need not be publication-polished but must be complete and reviewable.
3. **Provenance:** baseline/event delta, overlap with other submissions, material AI/tool/compute/credit disclosures, outside help, conflicts, and reproducibility limitations. Include prior knowledge and parent links for variations.
4. **Publication authority:** attribution approval and permission to release materials under the competition terms. Company entries must be authorized to represent the company and release the relevant work.

During the week, incomplete packets may receive a cure window of up to 12 hours, subject to the final cutoff. After cutoff, only administrative corrections—such as metadata or links to an artifact demonstrably fixed before the deadline—are allowed. New mathematics, repaired proofs, new theorems, or late formalization are not administrative cures. A narrowly documented transfer failure may use the pre-published fallback, not extra research time.

9 Review, governance, and appeals

9.1 Responsibilities and review sequence

The mathematical chair resolves classification/scoring disputes; difficulty assessors rate targets; stewards maintain status/families; expert judges review claims; the verification chair oversees formal/computational checks; operations maintain queues/deadlines/logs; reporting publishes accepted scope accurately. People may hold several roles but cannot review or approve their own work.

The review pipeline is:

1. Triage eligibility, packet, deadline, roster, provenance, and reproducibility.
2. Independently rerun the Hill/formal artifact and check its trust boundary.
3. Confirm admission, exact statement, modality, frozen difficulty, status, and family.
4. Obtain two qualified mathematical reviews of fidelity, novelty, logical scope, overlap, and attribution; reproduce computational components where required.
5. Record accepted claims, completeness/progress, D, b, m, p, F or the accepted M2 family/count as applicable, reasons, conflicts, and unresolved objections. Do not change difficulty to fit the result.
6. Resolve reviewer differences, seek written entrant clarification, and lock the accepted record after any appeal.

A complete solution requires the exact canonical target, no mathematical gap, permitted assumptions, passing checks, two qualified reviews, adequate literature checks, and resolved material objections. Otherwise label it a candidate or accepted partial, not a verified solution. Oral presentations and scheduling availability confer no points; written clarification and reruns suffice.

Maintain explicit states: draft, submitted, triaged, formal check failed/passed, under review, needs revision, accepted partial, accepted solution, rejected, contested, and final. Later corrections or retractions preserve the original decision and history. A failed formal check blocks mathematical acceptance and scoring; cure it within the permitted window or reject the claim. Provisional checking is not final acceptance.

9.2 Conflicts, confidentiality, and appeal

Disclose recent coauthorship, supervisory relationships, close collaboration, personal/financial interests, direct institutional interests, and involvement in the work. No one may be the sole reviewer of their student's or close collaborator's submission. The chair assigns unconflicted replacements wherever impartiality could reasonably be questioned. Only assigned reviewers and authorized organizers access private packets/deliberations. Rejected drafts remain private unless authors permit release; accepted scoring artifacts follow Section 10.

Appeals may identify mathematical error, misread claims, wrong prior/open status, inconsistent scoring, conflicts, or clerical mistakes—not stylistic preference. Appeal windows are 24 hours after a communicated ruling during the event and seven days after publication of the final record. An independent reviewer considers evidence and the chair records the outcome, recusing when implicated. Appeals cannot introduce mathematics missing at cutoff. Later advances may update the scientific record with attribution, not create a second award period.

10 Authorship, publication, and the three ledgers

10.1 Human attribution and public results

Human contributors receive full scholarly authorship for their new solutions, proofs, formalizations, code, exposition, and new formulations regardless of AI/tool use. Original problems and prior theorems retain their original attribution. AI systems, providers, sponsors, and organizers acquire no authorship or ownership by supplying tools/credits. Submission does not transfer contributors' existing copyright or require surrender of employer/third-party rights; entrants must hold the permissions needed to publish.

For accepted scoring work, entrants grant organizers nonexclusive worldwide permission to host, reproduce, verify, index, and include the submitted statement, proof/artifact, relevant code, disclosures, and approved text in the public record and bundled publication. Authors remain free to publish elsewhere, extend their work, and use other compatible licenses. Accepted scoring artifacts and necessary reproducibility material become public; indefinite embargo is incompatible with points.

Organizing institutions validate the stated review process and exact accepted claim, not broader unreviewed results. The final bundle lists authors, contributions, claims, artifact links, tools, and status. Unaccepted research leads require authors' permission and separate labeling. Competition proceedings do not substitute for external journal peer review.

10.2 Keep competition, mathematics, and system evaluation distinct

1. **Entrant competition:** cumulative S_T and biggest-accomplishment A_T within individual, team, and company classes, plus a separate M2 formalization-count leaderboard in each class. This measures accepted entrant contributions under these rules.

2. **Collective Mathematical Frontier:** deduplicated accepted original solutions, partial advances, known-mathematics formalizations, and variation results. Count a mathematical result once even if independent entrants each earn credit.
3. **Model and Harness Evaluation:** link attempts and accepted artifacts to disclosed systems, methods, and compute. Tool-use correlations do not establish causal model superiority or machine authorship; claims must match the evidence.

Keep an internal operations queue, a visibly provisional live view, and an authoritative accepted record. Public labels distinguish candidate; accepted formalized partial; verified M1/M3A solution; M2 formalization; M3B solution; verification contribution; and under-review/contested work. Use *open problem solved* only for accepted complete M1/M3A targets. Report candidates, accepted partials, and verified solutions separately, never as one undifferentiated count of results proved. Volume claims use deduplicated categories, not declarations, submissions, or trivial variants.

Public score records identify entrant/class, problem/family, modality, exact accepted claim, snapshot, Hill/Climb/commit, formal/review status, D, b, m, p, F where applicable, relevant disclosures, approved public review summary, latest review, total S, biggest-accomplishment A with its winning family, and the separate M2 count. Do not expose private review notes or unnecessary personal data.