SavchenkoSolutions Physics Benchmark — Scoring Specification

Release **v0.1** — Updated: **October 12, 2025**

Maintainer: **Aliaksandr (Alex) Melnichenka**
aliaksandr@melnichenka.com — melnichenka.com

Purpose & Scope

This document defines how submissions are scored for the SavchenkoSolutions Physics Benchmark (SSPB). It covers item types, normalization, partial credit, aggregation, error taxonomy, reliability statistics, and integrity rules. The submission format and schemas are defined in the companion *Schemas & Submission* PDF.

Item Types & Targets

Each problem is labeled with a `type`:

- **numeric**: final numeric value with required units.
- **symbolic**: closed-form expression (equivalence up to algebraic simplification).
- **proof**: short proof/derivation graded by rubric.

Some items include a `solution_outline` and `final_answer`. Multi-part items are split into atomic sub-items in the test set.

Normalization & Alternate Correct Answers

General

We first normalize whitespace, common symbols, and localized decimal/thousand separators. We also canonicalize units (SI preferred) and numeric formatting (scientific notation accepted).

Numeric alternates

If an item allows multiple equivalent forms (*e.g.*, $9.8\,\mathrm{m/s^2}$ vs. $980\,\mathrm{cm/s^2}$), the gold includes a set of acceptable values/units and tolerance windows. See Section 4.

Symbolic alternates

We accept mathematically equivalent expressions after canonicalization (commutativity/associativity, factoring, trig identities where declared). See Section 5.

Numeric Scoring

Let $x^\star$ be the gold value, $\hat{x}$ the predicted value. Two tolerances may be specified per item:

$$\text{absolute: } |\hat{x} - x^\star| \le \epsilon$$
$$\text{relative: } \frac{|\hat{x} - x^\star|}{\max(|x^\star|, \delta)} \le \tau$$

where δ avoids division by zero when $x^\star \approx 0$ (default $\delta = 10^{-9}$ in natural units).

Units. Units must match exactly after canonicalization (e.g., $\mathrm{N\,m} \equiv \mathrm{J}$). Answers with correct magnitude but wrong/missing units are marked incorrect unless the item explicitly flags "unitless".

Significant figures. We do *not* penalize sig-figs if the value falls within tolerance, but we compute a *sig-fig sanity* flag for diagnostics.

Score.

$$s_{\text{numeric}}(\hat{x}) = \begin{cases} 1 & \text{if within tolerance and units correct} \\ 0 & \text{otherwise} \end{cases}$$

An optional partial credit (+0.5) may be awarded if the method is correct (from `reasoning`) but a minor algebra/rounding slip exceeds the narrow tolerance; this is used only when the item metadata enables `allow_method_partial=true`.

Symbolic Scoring

We canonicalize both gold and prediction using a CAS pipeline (SymPy):

1. Parse to AST; standardize variables/constants.
2. Simplify using algebraic rules declared per item (e.g., trig identities allowed?).
3. Optionally expand/factor to a canonical polynomial form.

If $\texttt{simplify}(\hat{E} - E^{\star}) = 0$ under declared domains/assumptions, the answer is correct.

Score.

$$s_{\text{symbolic}}(\hat{E}) = \begin{cases} 1 & \text{if proven equivalent under allowed rules} \\ 0.5 & \text{if structurally correct form but minor algebraic slip} \\ 0 & \text{otherwise} \end{cases}$$

The 0.5 case is triggered by CAS returning a low-complexity residual or editor-marked near-miss patterns.

Proof Scoring (Rubric)

Proof items are graded by rubric with weighted criteria:

Criterion	Description	Weight
Correct statement of goal	Clearly states what must be shown	0.10
Key lemmas/claims	Identifies correct intermediate results	0.25
Logical flow	Steps follow with valid inferences	0.30
Handling of edge cases/assumptions	Domains, boundary conditions, units	0.15
Clarity/structure	Readable, minimal gaps, correct notation	0.10
Conclusion	Explicitly closes the argument	0.10

Each criterion is scored in $\{0, 0.5, 1\}$ and multiplied by the weight. The proof score $s_{\text{proof}} \in [0, 1]$ is the weighted sum. We sample proofs for dual grading; Cohen's κ is reported.

Aggregation & Metrics

Let S be the set of test items. We report:

- **Overall accuracy:** $\frac{1}{|S|}\sum s_i$ (numeric/symbolic treated as 0/1; proof is s_{proof}).
- **Per-type accuracy:** numeric, symbolic, proof.
- **Per-topic accuracy:** mechanics, E&M, etc.
- **RU/EN split:** language-wise metrics.
- **Units compliance rate:** for numeric.
- **Error taxonomy rates:** Section 9.

We provide 95% CIs via nonparametric bootstrap (10,000 resamples).

Calibration & Significance

We test differences between systems using paired bootstrap on per-item scores. We flag differences as statistically significant at $\alpha = 0.05$ after Holm–Bonferroni correction for multiple comparisons.

Error Taxonomy

Each incorrect item is tagged (automatically, then optionally audited) with:

- Units/dimensions error
- Algebraic manipulation error
- Physical modeling error (wrong regime/assumption)
- Boundary/initial conditions mishandled
- Problem misread / spec noncompliance
- Hallucinated constant/law

Runtime Budget & Integrity

- **Time:** Soft guidance $\leq 60\,\text{s}$ per item median; report median latency.
- **External calls:** Allowed tools must be declared; web search on the exact item text is prohibited.
- **Caching:** Disclose any retrieval of prior SSPB items; test split is unpublished before release.
- **Human-in-the-loop:** Not allowed for official runs.

Submission Validation

Payload must conform to `predictions.schema.json`. Minimal example:

```
{
  "run_id": "team-YYYYMMDD",
  "predictions": [
    {"problem_id": "SS-EN-01234", "answer": "9.81 m/s^2", "reasoning":
       "..." }
  ]
}
```

Audits, Appeals & Versioning

We audit random 5–10% of items per run. Appeals on specific items are accepted within 14 days; please cite item ID and rationale. Scoring changes trigger a minor version bump; gold updates trigger a new test release with a new hash.

Contact

Aliaksandr (Alex) Melnichenka
aliaksandr@melnichenka.com — melnichenka.com