# SavchenkoSolutions Physics Benchmark — Dataset Card

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

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

## Executive Summary

**Purpose.** Provide a clean, license-clear subset of SavchenkoSolutions to evaluate reasoning of LLMs/solvers on physics problems (RU/EN), with scoring and an error taxonomy.
**Pilot model.** Labs run inference via batch API; we return metrics and a short public *aggregate* note with baselines. No NDA, no proprietary code/data exchange, education-only.

## Intended Use

- **Primary:** Research/education evaluation of physics reasoning for models/agents.
- **Not for:** Commercial benchmarking commitments, model comparison marketing, redistribution of raw items.

## Provenance & Composition

The benchmark draws from the SavchenkoSolutions archive; we include only items with clear educational licenses or our own authored/edited content.

| | |
|---|---|
| **Items** | ∼5,000 physics problems with structured fields: statement, hints, solution outline, final answer. |
| **Languages** | Russian and English (parallel where available). |
| **Topics** | Algebra, mechanics, thermodynamics, E&M, optics, modern physics. |
| **Formats** | Numeric (with units/tolerances), symbolic (closed-form), proof-style (claims/lemmas). |
| **Metadata** | Topic, difficulty (1–5), prerequisites, tags, expected units, type. |
| **Splits** | Suggested: train/dev/test; disjoint by item ID and near-duplicates (see Quality). |

## Licensing & Access

- **License:** Education/research-only per SavchenkoSolutions Terms. Redistribution of raw items is prohibited.
- **Attribution:** Cite this dataset card and the SavchenkoSolutions site; include link to the public pilot report if used.
- **Compliance:** Open, non-exclusive pilot avoids export-control/IP entanglements; no NDAs.

## Quality & Curation

- **Deduplication:** Hash-based near-duplicate removal and manual spot checks.

- **Units/dimensions:** Automated consistency checks for numeric items; unit canonicalization.
- **Language parity:** RU/EN parallelism flagged; parity sampled for drift.
- **Difficulty calibration:** Heuristic + editor consensus; will be refined with pilot feedback.

## Ethical Considerations

- **Academic integrity:** Items are for evaluation/education; not for assisting live course assessments.
- **Privacy:** No personal data included.
- **Transparency:** Public, aggregate reporting only; labs may remain unnamed in public notes.

## Suggested Splits

- **Train:** up to 60% (if you fine-tune on similar content).
- **Dev:** ∼20% (hyperparameters/evaluation sanity).
- **Test:** ∼20% (held-out). The pilot uses a specific test release pinned by hash.

## Scoring Overview (Summary)

Detailed rubric in the *Schemas & Submission* document.

- **Numeric:** tolerance window + units check + sig-fig sanity.
- **Symbolic:** canonicalization with SymPy where viable; exact/normalized match otherwise.
- **Proof-style:** rubric (claims/lemmas, key step coverage); sampled human audit.
- **Partial credit:** +1 correct final; +0.5 correct method w/ algebra slip; +0.2 correct setup only.

## Error Taxonomy (Pilot)

- Units/dimensions; algebraic; physical model; boundary/initial conditions; misread; hallucination.

## Submission & API (Pilot)

- **Mode A (preferred):** Upload predictions JSON (signed URL).
- **Mode B:** Fetch test slice via API, return within 72h.
- **Security:** Per-team keys; rate limits; logs retained for audit.

## Public Reporting

We publish a short, non-comparative note: methodology, baselines, *aggregate* metrics and error breakdowns. No prompts, weights, or per-item outputs.

## Contact

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