SavchenkoSolutions Physics Benchmark — Baselines & Reference Results

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

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

Purpose

This document reports *reference* baselines to anchor the benchmark. These are illustrative, not endorsements. Settings and prompts are fixed for reproducibility; see the Repro Checklist.

Evaluation Protocol (Summary)

- Test split: `SSPB-test-v0.1` (hash: `REPLACE_HASH`).
- Metrics: overall accuracy, per-type, per-topic, RU/EN. 95% CIs via bootstrap.
- Runtime: median per-item latency measured wall-clock.

Hardware & Runtime Budget

Compute	Single A100-40GB (or CPU for non-LLM baselines)
Max per-item time (soft)	60 s median
Tool access	SymPy & NumPy permitted; no internet retrieval

Baseline Systems

B0: Random-Guess + Units Coinflip

Trivial baseline to calibrate floor; picks from multiple-choice (if any) or random numeric within coarse range; units at 50% coinflip.

B1: Dimensional Analysis Heuristic

Extracts quantities; solves by dimensional homogeneity and common constants; no equation solving.

B2: CAS Solver (Symbolic-first)

Parses statement patterns; uses SymPy to set up equations for single-unknown numeric items and closed-form algebra for symbolic items. No multi-step physical reasoning.

B3: LLM Zero-Shot (Text-Only)

Large language model with a concise prompt; greedy decoding or temperature $t = 0.2$; no tools.

B4: LLM + Chain-of-Thought (No Tools)

Same as B3 but encourages step-by-step derivation; final answer extracted after a `<final>` tag.

B5: Toolformer (LLM + SymPy/Units)

Agent uses a restricted toolset (SymPy, unit checker). The prompt enforces: compute in tools, return clean final with units.

B6: Agentic Solver (LLM + Planner + Tools)

Lightweight planner decomposes into subgoals (free-body diagram, laws, solve, units). Enforces a 5-step template and unit validation before emitting final.

Prompting & Settings (LLM Baselines)

- Decoding: temperature 0.2, top-p 0.95, max tokens 512.
- Stop: `<final>` tag.
- RU items: bilingual prompt; model may answer in English but must preserve symbols/units.

Reference Results (Illustrative Placeholders)

Replace with your actual runs before public release.

Overall

System	Overall	Numeric	Symbolic	Proof	RU	EN
B0 Random	0.06	0.05	0.02	0.10	0.05	0.07
B1 DimAnal	0.18	0.26	0.05	0.08	0.15	0.21
B2 CAS	0.34	0.41	0.52	0.07	0.28	0.39
B3 LLM ZS	0.37	0.32	0.44	0.31	0.33	0.40
B4 LLM CoT	0.45	0.39	0.50	0.46	0.41	0.49
B5 LLM+Tools	0.55	0.63	0.57	0.40	0.50	0.59
B6 Agentic	0.60	0.68	0.60	0.50	0.56	0.63

Note: Values above are placeholders. Insert true means and 95% CIs.

Per-Topic Breakdown (Example)

System	Mech	Thermo	E&M	Optics	Modern
B2 CAS	0.46	0.28	0.40	0.35	0.18
B5 LLM+Tools	0.66	0.54	0.60	0.58	0.48
B6 Agentic	0.71	0.61	0.66	0.63	0.55

Ablations

- **No units check** (B5†): expect +2–3% raw but higher error taxonomy rate for units.
- **No planner** (B6→B5): drop in proof rubric and multi-step numeric.
- **Higher temperature** ($t = 0.8$): more diverse but lower numeric accuracy.

Error Analysis (Snapshot)

System	Units	Algebra	Model	BC/IC	Halluc.
B3 LLM ZS	18%	22%	28%	8%	24%
B5 LLM+Tools	6%	15%	20%	7%	10%
B6 Agentic	5%	12%	17%	6%	8%

(% of incorrect items assigned the tag; multi-tags allowed.)

Cost & Latency

System	Median (s)	P90 (s)	Notes
B2 CAS	0.8	2.1	Fast; fails on multi-step modeling
B4 LLM CoT	6.2	15.4	No tools; longer traces
B5 LLM+Tools	8.5	18.7	Tool calls dominate
B6 Agentic	10.3	22.0	Planner adds overhead

Repro Checklist

- Commit hash of test set: `REPLACE_HASH`
- Exact prompts (appendix or linked repo)
- Decoding parameters
- Tool versions: SymPy, NumPy, Python
- Random seeds and batching strategy
- Hardware description

Notes & Disclaimers

Baselines are for orientation only. Do not cite as state-of-the-art. Any public comparison to named third-party systems requires their consent and ours.

Contact

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