

PROTECTION-FIRST BENCHMARK BRIEF

If PassLab passes it, submit it.

PassLab is built around a different finish line. A lower detector-facing score is not enough. A delivered rewrite must also preserve the argument, evidence, citations, grammar, and whole-document usability.

8/20	BypassAI outputs with an apparent ≥ 10 pp win	0/20	BypassAI outputs that survived the full release gate
21.14	StealthWriter mean home-score drop in expanded64	3.33	mean drop across PassLab's broader equal-family view

Others stop when the score drops. PassLab stops when the paper is ready.

This brief documents the current experiment inventory, competitor run methods, dates, result definitions, negative evidence, and limitations behind that product direction. It is an evidence brief, not a completed world ranking.

DOCUMENT	SCOPE	STATUS
Evidence Brief v0.1	Humanizer and detector-direction evidence through 13 Jul 2026	Public methodology preview
Claim boundary	Protection-first delivery architecture	Supported
World-best ranking	Representative market-wide superiority	Not yet established

01 / THE CATEGORY FAILURE

Why a lower score can still leave unusable writing

A score-only humanizer can appear successful while creating a second job for the user. The number falls, but meaning shifts, evidence disappears, citations move, grammar degrades, or the output is truncated. Repairing those defects can then restore the same detector-facing signals the user was trying to reduce.

STEP	WHAT THE USER SEES	WHAT THE SCORE MISSES
1. Score drops	A more favorable detector-facing number	Whether the output still says the same thing
2. Text breaks	Awkward or superficially varied language	Claim loss, entity loss, citation damage, truncation
3. User repairs	Meaning and readability return	The user is now doing the quality work
4. Score rebounds	The rewritten paper needs another pass	The product has created a retry loop

PassLab's unit of success

A rewrite counts only when detector-facing movement and deliverability survive together. Internally this is framed as qualified meaningful-drop coverage: meaningful risk movement, hard protection passed, usable language, and no whole-document regression.

DETECT	TARGET	PROTECT	RELEASE
Read multiple local risk directions.	Choose the risks that matter.	Reject damaged candidates.	Deliver only a cleared rewrite.

Every PassLab rewrite is submission-ready - or the original stays.

02 / EXPERIMENT INVENTORY

What was tested, and when

The rows below are deliberately not collapsed into one leaderboard. The experiments answer different questions, use different evidence conditions, and carry different completion states.

SYSTEM / TRACK	SCOPE	DATE	OBSERVED RESULT	STATUS
Phrasly	30 unique sources; 68 retained candidates	12 Jul 2026	Real raw reductions, especially on shorter text; recurring grammar, citation, claim, truncation, and generation-residue defects.	Characterization complete
BypassAI	20 sources; 21 retained candidates	12 Jul 2026	Large raw drops were possible. All eight candidates with ≥ 10 pp movement failed the combined manual quality/release audit; qualified coverage was 0/20.	Characterization complete
WriteHuman	20 common sources; up to five candidates	12 Jul 2026	Targeted selection raised raw ≥ 10 pp coverage from 10/20 to 13/20 before full protection penalties.	Blind protection audit pending
StealthWriter	952 historical Tier A pairs; 64 deduplicated matched pairs	11 Jul 2026	Mean self-reported drop was 21.14 pp in expanded64; broader PassLab equal-family movement averaged 3.33 pp. No pair reached a 20 pp away drop.	Human blind review pending
PassLab x Pangram	97 unique Pangram observations; 114 logical rows	13 Jul 2026	In a 15-text delivery slice, two moved 100 to 0 and three moved 0 to 100. The audit separated generator and selector failures.	Diagnostic complete
Grammarly sentence protocol	327 of 327 completed; 966 of 966 sentence signals aligned	14 Jul 2026	A corrected waiting and extraction protocol materially changed 33 competitor records and invalidated the older capture method.	Protocol closeout complete
Scribbr / QuillBot	181 of 181 completed competitor runs	14 Jul 2026	Platform outputs varied sharply by product and sample. Detector success did not establish meaning protection or overall usability.	Run closeout complete

Important denominator note: 952 is an attempt-level historical archive, not 952 independent qualified deliveries. The 64-pair StealthWriter slice is deduplicated by source and parent-lineage proxy. The Pangram count separates unique observations from logical rows that reuse identical canonical text.

03 / RUN PROTOCOL

Why the experiments have weight

The benchmark method is designed to make favorable cherry-picking harder and failure evidence harder to hide.

01	Same-source comparison	A benchmark ID carries the same frozen source text, line breaks, and source hash across systems.
02	Public mode separation	Default and strongest public modes are separate runs. A stronger mode cannot silently replace a failed default run.
03	One official output per lane	Manual cleanup is not presented as vendor output. Retrying until a favorable result appears is not the standard lane.
04	Failures stay in the denominator	No-output events, truncation, refusals, errors, and unusable text remain part of product usability.
05	Home, away, and neutral scores	Vendor self-scores are home evidence. PassLab local families are away evidence. Human blind review is a separate neutral layer.
06	Protection before promotion	Meaning, claims, sources, citations, names, numbers, grammar, and length behavior are checked before a rewrite becomes a qualified result.
07	Provenance and versioning	Run IDs, dates, modes, input and output hashes, screenshots or raw snapshots, and product state are retained where available.
08	Observed, inferred, unknown	Reports separate direct observations from interpretations and from questions the current evidence cannot answer.

04 / FINDINGS

The evidence supports a different finish line

The strongest recurring result is not that every competitor is weak. It is that detector movement and deliverability are separate axes.

FINDING	EVIDENCE	PRODUCT IMPLICATION
Large drops can be purchased with damage	BypassAI produced large local movements, while all eight >=10 pp cases failed the combined manual release audit.	Risk movement cannot be the only reward or public win condition.
A vendor's home score can overstate cross-judge movement	In expanded64, StealthWriter's mean home drop was 21.14 pp versus 3.33 pp under a broader equal-family away aggregation.	Show multiple risk directions and keep vendor self-scores separate.
Candidate choice matters	WriteHuman's best-of-five selection improved raw coverage; PassLab's Pangram audit found cases where safer candidates existed but the selector missed them.	Generate alternatives, then apply hard protection before selection.
Low-risk input needs protection too	The Pangram diagnostic included three 0 to 100 regressions in the 15-text delivery slice.	No-regression is a delivery requirement, not optional polish.
Capture methods can change the apparent result	The corrected Grammarly sentence protocol changed 33 competitor records and invalidated the older waiting approach.	Freeze extraction protocol and retain raw snapshots.

What this supports today

PassLab can substantiate a protection-first product architecture and a stricter definition of a usable result. It can show that its benchmark work includes named competitors, real run dates, failed outputs, corrected protocols, and negative results about PassLab itself.

What this does not support today

The current packet does not establish that PassLab is the best humanizer in the world. It does not yet contain a current, representative, same-source, fully blinded world ranking with complete semantic and citation adjudication for every major competitor.

05 / RELEASE DEFINITION AND LIMITATIONS

What 'Passed' must mean

The homepage promise depends on a strict product contract. A rewrite should display Ready to submit only after the release path has cleared the required protection checks. If no candidate is safer than the source, the final text must remain the original rather than forcing a rewrite.

RELEASE CHECK	PASS CONDITION	FAILURE BEHAVIOR
Meaning and claim direction	The candidate preserves the source argument and does not introduce a material claim shift.	Reject candidate
Facts, names, and numbers	Required entities and numeric claims survive without unsupported additions or deletions.	Reject candidate
Sources and citations	Citation presence and source-linked wording remain intact under the applicable scope.	Reject candidate or restore source
Grammar and completeness	The output is complete, readable, and free from terminal corruption or leakage.	Reject candidate
Whole-document no-regression	The final candidate is not worse than the original under the release policy.	Keep original

Known limitations

1. Local risk directions are estimates, not official detector verdicts. Detector versions and score surfaces can change.
2. Several competitor tracks remain interim because neutral blind semantic and citation review is incomplete.
3. Product capability and selector reliability can change after model, routing, or policy updates. Public claims must carry version and date context.
4. Academic submission rules vary. 'Ready to submit' describes the PassLab release checks, not institutional permission to submit work the user did not author.
5. The evidence is strongest as a transparent product-design and comparison record. A market-wide superiority claim requires a new frozen, matched, representative benchmark.

NEXT EVIDENCE GATE

Freeze a representative same-source set, run current public competitor versions, finish blind protection adjudication, publish the scoring contract, and retain all failures before promoting a global ranking.

Humanize once. Submit once.

PassLab Evidence Brief v0.1 / prepared 6 Aug 2026 / evidence current through 14 Jul 2026