AI ENABLEMENT LABS · EVIDENCE BRIEF

The research behind a verified workflow

Seven claims we make, the decisions they justify, and their limits.

This brief sets out the published evidence underpinning our method. It is written to be cited in your own governance, board, and assurance documents. We organise it by *claim* rather than by paper: each finding is paired with the decision it justifies in our method and, deliberately, with its limits. Where evidence is provisional or contested, we say so — overstating the research would defeat the point of having it.

CLAIM 1

For most knowledge work, the scarce human skill is shifting from producing to verifying.

Why it shapes our method. We treat Verify as a distinct, resourced stage of the workflow rather than an afterthought.

The evidence. The “generation–verification gap” is a named, studied problem: models often produce a correct answer yet cannot reliably distinguish their correct answers from their wrong ones. Closing the gap takes deliberate machinery — combining several independent, imperfect checkers can approach the accuracy of a perfect verifier far more cheaply than scaling the model.

Limits and caveats. Verifying is cheaper than generating in principle, but it is neither automatic nor free. A model checking its own work in the same context is unreliable; the gains come from independent and combined checks. As models improve, the errors that remain get subtler — which raises the value of structured verification rather than removing the need for it.

Primary sources. Saad-Falcon et al. (2025), *Shrinking the Generation–Verification Gap with Weak Verifiers* (Weaver), NeurIPS 2025. arXiv:2506.18203.

CLAIM 2

AI raises the floor of knowledge work more than the ceiling.

Why it shapes our method. It underwrites our promise to raise both floor and ceiling, our focus on less-experienced staff, and the operating modes that let a beginner ship to the same standard as a principal.

The evidence. In a pre-registered field experiment with 758 Boston Consulting Group consultants using GPT-4, those with AI were more productive and produced higher-quality work on tasks within AI’s capability — and the largest relative gains went to the lower-performing participants. The skill distribution compressed upward.

Limits and caveats. Measured on consulting tasks with a GPT-4-era model; the exact magnitudes are specific to that setting and will move as tools and tasks change. “Higher average quality” is not “uniformly safe” — see Claim 3.

Primary sources. Dell’Acqua, McFowland, Mollick, Lakhani et al. (2023), *Navigating the Jagged Technological Frontier*, HBS Working Paper 24-013 (forthcoming, Organization Science). SSRN 4573321.

CLAIM 3

On tasks outside AI's "frontier," output gets worse — and people don't reliably notice.

Why it shapes our method. It is the reason for the Frame stage: decide what "good" looks like and where the risk sits before seeing a persuasive draft. It is also why we risk-tier work and make checks mandatory.

The evidence. In the same experiment, on a task designed to sit just outside AI's frontier, consultants using AI were around 19 percentage points *less* likely to reach the correct answer than those without it — they followed fluent but wrong output. The frontier is "jagged": near-identical-looking tasks can fall on opposite sides of it.

Limits and caveats. The frontier moves as models improve, so any fixed list of "inside" and "outside" tasks dates quickly. The durable lesson is the shape of the risk — uneven and hard to see — not a specific boundary.

Primary sources. Dell'Acqua et al. (2023), as above.

CLAIM 4

Domain expertise alone does not protect against AI's failures.

Why it shapes our method. It is the basis of our two-axis diagnostic — subject knowledge and AI-failure literacy are independent skills — and of training failure literacy even for senior specialists (the "fooled expert").

The evidence. The consultants above were experienced professionals and still followed AI into wrong answers on out-of-frontier work. Separately, evaluation research shows that judgement made without an independent reference is unreliable and driven by surface features such as fluency and confidence — exactly what slips past someone who knows the subject but isn't watching for it.

Limits and caveats. This concerns a specific failure mode — over-trusting fluent output — not a claim that expertise is unimportant. Expertise is necessary for good review; it is simply not sufficient on its own.

Primary sources. Dell'Acqua et al. (2023); reference-grounded evaluation literature (Claim 7).

CLAIM 5

Heavy reliance on AI is associated with weaker critical thinking; skills and method buffer it.

Why it shapes our method. It is why we build verification capability deliberately and measure success as people getting sharper, not more dependent — the method is designed to keep humans thinking, not to think for them.

The evidence. A mixed-methods study of 666 participants found a significant negative correlation between frequent AI-tool use and critical-thinking scores, mediated by increased cognitive offloading; higher educational attainment buffered the effect.

Limits and caveats. Correlational and partly self-reported — it shows association, not proof that AI use causes decline — and the paper has since carried a published correction. We treat it as a credible warning that informs design, not settled causation.

Primary sources. Gerlich (2025), *AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking*, Societies 15(1):6. doi:10.3390/soc15010006.

CLAIM 6

AI output is measurably prompt-sensitive and unstable, so how a tool is used dominates the result.

Why it shapes our method. It is why buying the tool is never enough, why we standardise prompts, context, and conventions, and why the diagnostic treats operator skill as decisive.

The evidence. A psychometric framework applied to 18 models found that measurable properties of their output are reliable and valid only under specific prompting configurations, and stronger for larger, instruction-tuned models. The same model behaves differently depending on how it is prompted — so the configuration and the operator, not just the model, decide the outcome.

Limits and caveats. This particular work concerns psychological and behavioural traits, so we extrapolate to general output quality with care. The transferable point is narrower but solid: AI output is configuration-dependent and needs validating.

Primary sources. Serapio-García et al. (2025), *A psychometric framework for evaluating and shaping personality traits in large language models*, Nature Machine Intelligence 7:1954–1968. doi:10.1038/s42256-025-01115-6. Supporting: arXiv:2505.08245; arXiv:2507.04491.

CLAIM 7

Checking work against an independent reference beats unreferenced judgement — and helps weaker reviewers most.

Why it shapes our method. It is the empirical case for validation touchstones — our core IP and recurring service — and for our first design rule: the reference must be independent of the thing it checks.

The evidence. Reference-grounded evaluation reliably outperforms reference-free evaluation, with the largest gains for weaker checkers (one method lifted a small model's accuracy by roughly 17 points) and lower variance between judges; reference-based checking against a gold answer reaches near-human agreement on factual tasks. Conversely, ungrounded judgement can fall toward chance, and a shared-but-empty template can manufacture false agreement.

Limits and caveats. Touchstones only help if they are genuinely independent of the model being checked and encode real ground truth. A same-model self-check, or a shared template with no substance behind it, produces sycophantic or illusory agreement, not validation. Touchstones also decay, and must be owned and refreshed.

Primary sources. *References Improve LLM Alignment in Non-Verifiable Domains* (RefEval, 2026), arXiv:2602.16802; *Beyond the Illusion of Consensus* (2026), arXiv:2603.11027; DAFE (Badshah & Sajjad, 2025).

Currency and ownership

This brief is itself an touchstone, and touchstones decay — so we hold it to our own rule. It has a named owner, a review date, and a refresh cycle. **Last reviewed: June 2026.** If a finding here is overstated, or new research complicates a claim, that triggers a revision. One canonical evidence base feeds this brief, the website, and the deck, so a single update propagates everywhere.

References

1 Dell’Acqua, F., McFowland III, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. Harvard Business School Working Paper 24-013 (forthcoming, Organization Science). https://ssrn.com/abstract=4573321

2 Gerlich, M. (2025). AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. Societies, 15(1), 6. https://doi.org/10.3390/soc15010006 (see published correction, Societies 15(9), 252).

3 Saad-Falcon, J., Buchanan, E. K., Chen, M. F., et al. (2025). Shrinking the Generation–Verification Gap with Weak Verifiers (Weaver). NeurIPS 2025. https://arxiv.org/abs/2506.18203

4 Serapio-García, G., Safdari, M., et al. (2025). A psychometric framework for evaluating and shaping personality traits in large language models. Nature Machine Intelligence, 7, 1954–1968. https://doi.org/10.1038/s42256-025-01115-6

5 LLM Psychometrics: A Systematic Review of Evaluation, Validation, and Enhancement (2025). https://arxiv.org/abs/2505.08245

6 A validity-guided workflow for robust large language model research in psychology (2025). https://arxiv.org/abs/2507.04491

7 References Improve LLM Alignment in Non-Verifiable Domains (RefEval) (2026). https://arxiv.org/abs/2602.16802

8 Beyond the Illusion of Consensus: From Surface Heuristics to Knowledge-Grounded Evaluation in LLM-as-a-Judge (2026). https://arxiv.org/abs/2603.11027