AidGPT

Verification-First

Commissioning AI-assisted analysis you can defend: a working protocol from two market assessments

AidGPT Guidance Paper MI-GUIDE-002 · Thomas Byrnes, MarketImpact Digital Solutions Ltd · Version 1.0, 17 August 2026 · Status: Guidance paper. Prepared ahead of the Markets in Crises Community of Practice fireside "AI on the Ground", 19 August 2026 · Canonical URL: aidgpt.org/AidGPT-MI-GUIDE-002-Verification-First.pdf · Correspondence: thomas@marketimpact.org ·

Abstract

AI has collapsed the cost of producing analysis. It has not collapsed the cost of verifying it, and the gap between the two is now the defining quality problem in humanitarian assessment. This paper sets out a working protocol for AI-assisted market analysis, developed and stress-tested in March 2026 on two desk-based market assessments, preparatory work for a Mercy Corps analysis series published in May 2026: a full Emergency Market Mapping and Analysis (EMMA) for Somalia, drafted from a blank page inside a prepared workspace, and an adversarial audit of an existing Pakistan draft. The protocol treats verification as half the job, in sequence: the workspace is provisioned before drafting starts, every claim carries an evidence grade as it is written, an adversarial audit runs as a mandatory gate before anything is shared, a robustness review hunts for what the audit cannot see, and the analyst owns every judgement. The paper documents nine failure modes that occurred in the two cases, each paired with the control that caught it, including a plausible geography error, a circular citation, and a validated fact that went stale within ten days. It closes with ten requirements that any organisation commissioning AI-assisted analysis can write into its terms of reference, so that speed gains arrive with their warranty attached. The workflow's value is not that the AI is right. It is that the workflow makes being wrong visible, tiered, and fixable before anyone else sees it.

Keywords: AI-assisted analysis, verification, market assessment, EMMA, evidence grading, adversarial audit, humanitarian analysis, commissioning standards, quality assurance

1. The problem: production is cheap, trust is not

A desk-based market assessment that took one to two weeks to research and draft can now be produced, audited twice, and rebuilt in a single working session. It happened, twice, in the cases this paper documents: a ten-section EMMA for Somalia with fifty-one cited sources built in a single session, and a one-hundred-plus data point Pakistan draft put through a structured audit that found two critical problems that had survived drafting and validation.

The speed is real. So is the danger. The same tools that produce evidence-dense analysis quickly also produce confident, plausible, well-formatted analysis that is wrong, and they produce both at the same speed with the same fluency. What separates the two is the workflow around the model, and specifically whether verification is budgeted as a first-class phase or treated as something the reader will do.

This paper documents a workflow in which roughly half of every session is spent on audit, robustness review, and rebuilding. Anyone selling you "the assessment in an hour" is selling the unaudited draft, which is the dangerous product. The draft in an hour exists. The warranty takes the second hour, and the second hour is the one that matters.

2. The evidence base

The protocol was developed and tested on two cases run in March 2026 as preparatory work for a real commission: the analysis MarketImpact carried out for Mercy Corps on the Iran war's secondary economic impacts, published, after full human verification, as From Hormuz to the Frontlines of Hunger, a six-context series including Somalia (mercycorps.org/research-resources/hormuz-to-hunger). That series is the finished product. This paper shares the internal method behind its working drafts.

Case 1, Somalia, built from scratch. A full ten-section desk-based EMMA produced from a blank page inside a prepared workspace (Phase 0 below). First draft, then adversarial audit, which found eighteen issues across seven categories, three of them critical. Then a multi-angle robustness review, which found seven entire analytical dimensions the audit had not flagged. Then a version three rebuild. The final document carried more than twenty-eight validated data points, fifty-one sources, and fifteen explicitly flagged gaps, each with a specified key-informant action.

Case 2, Pakistan, audited. An earlier AI-drafted assessment from the same programme, carrying one hundred and one validated data points, put through the same audit methodology as a quality gate. The audit found two critical problems. A fact had gone stale: a fertiliser plant reported as shut down had resumed operations between drafting and review. And a structural context was missing: the country had held a gas surplus weeks before the crisis, which reframed the entire supply narrative. Thirteen further prioritised fixes followed.

Two results from these cases anchor everything that follows. The Somalia audit, hunting specifically for fabricated claims, found zero hallucinations in the AI-drafted text, a result that came from the tagging discipline described below. And the Pakistan audit showed the gate earning its keep on finished work: a draft with over a hundred validated data points, built earlier under the same method, still carried two critical problems, one gone stale after drafting, and one that checking the stated facts could never find, because the fact was missing. A draft can pass validation and still fail verification, and the audit is the pass that tells them apart, provided the person running it can read what it returns.

The Somalia working draft is published with this paper as a case exhibit, tags, gap register and evidence summary intact: aidgpt.org/AidGPT-MI-GUIDE-002-Somalia-Case-Exhibit.pdf

3. The protocol

Seven phases, and a step before any of them. Roughly half the session is drafting, half is verification.

Phase 0, provision the workspace. The protocol runs inside a prepared workspace, never a blank chat. Load the governing methodology into the project as its standing reference material, and give the AI a clear brief: the role it is playing, the methodology that governs the work, and the rules it cannot break, the tagging discipline first among them. These tools perform best under clear instructions, and market analysis is unusually well provisioned, because EMMA already specifies in detail what a good assessment contains. In the cases here, the full EMMA toolkit and guidance sat in the project throughout, so the model never had to invent the structure of the document. The framework disciplines the shape. It does not verify the content, which is what the rest of the protocol exists to do.

Phase 1, frame before searching. Start with a transmission hypothesis, not a blank question. The hypothesis gives every search a target and gives the audit something to attack. A hypothesis you cannot attack is not a hypothesis. The analyst writes or approves it. The AI does not invent the analytical frame.

Phase 2, data assembly, local-language first. The single highest-value sourcing lesson, confirmed in both cases: local and regional media first. International wires give the direction of change. Local outlets give the numbers, the baselines, and the texture, in one case the actual pump prices and the arrest at a fuel protest that opened the domestic political dimension. Global upstream data is validated once and reused across every geography.

Phase 3, draft directly into the final structure, every claim tagged. Four grades, applied inline from the first sentence:

Tag	Meaning
VALIDATED	Two or more independent sources
SINGLE SOURCE	One credible source, flagged for key-informant verification
INFERRED	Analytical inference, logic chain documented
GAP	No current evidence, follow-up action specified

The AI is not permitted to write a claim without a tag, and a tag forces a source. This is the structural control against hallucination, and it is what produced the zero-hallucination audit result.

Phase 4, adversarial audit as a mandatory gate. Fixed categories: unsourced claims, hallucination checks, overstatement, missing caveats, structural completeness, cross-reference against the wider evidence base, and arithmetic. Every issue gets a severity tier and a specific fix. Critical and high fixes are applied before anything is shared. Asking the drafting pass to critique its own output does not count. That is self-critique, not independent proof, and the audit is a separate pass for exactly that reason.

Phase 5, multi-angle robustness review. The audit checks what is on the page. The robustness review asks what should be on the page. In the Somalia case the audit found eighteen issues, and the robustness review then found seven missing analytical dimensions the audit had no way to see, including a healthcare-access transmission chain and a religious-calendar collision with peak livestock exports. They are different questions, and one pass of either is not enough.

Phase 6, rebuild, revalidate, count. Apply fixes, rebuild, and update the evidence counts. Gap counts going up is a feature: it means the review found real unknowns and named them instead of papering over them. Validated counts going down is a red flag on the earlier draft, and the reason should be investigated.

Phase 7, human review and key-informant handoff. The analyst reviews the final draft, owns every analytical judgement in it, and takes the gap list into interview planning. A desk assessment's job is to make the field verification surgical: every gap carries a named informant type and a specific question.

4. Nine failure modes, and the control that caught each

Every failure below occurred in the two cases. Each was caught by a specific, repeatable control. Teach the failure and the control together, because a rule without its failure attached is advice, and advice does not survive a deadline.

Failure mode	Live example	Control that caught it
Confident overstatement	"The only geography with every channel active", when the project's own data showed two others at the same level	Overstatement check against the project's own evidence
Plausible geography error	A shipping route framed as dependent on a chokepoint it does not transit	Mechanism check: trace the physical route
Circular citation	A draft citing the project's own earlier output as a source	No-self-citation rule in the unsourced-claims check
Staleness	A plant shutdown reported as current after operations had resumed	Recency re-search on every current-status claim in the sign-off week
Missing structural context	A gas surplus and a prior allocation decision absent from a supply narrative built entirely of true facts	Contrarian search: what is the strongest evidence against this framing
Figure conflation	Two fuel products merged into one price series	Data verification pass: split commodities, source each baseline
Misattributed secondary figures	An export value attached to the wrong year by an aggregator	Trace to the primary source, never settle on the aggregator
Single source presented confidently	A price baseline resting on one report	Tagging discipline plus targeted re-search
Unmetriced superlatives	"Most import-dependent" without a measure	Every superlative names its metric or is softened

Two of these deserve emphasis, because they are the ones that hurt practitioners who skip the controls.

Staleness is the silent killer. In a fast-moving crisis, a claim can be validated, multi-sourced, and wrong within ten days. The control is mechanical: any claim about current operational status gets a fresh search dated within the sign-off week, regardless of how well sourced it was at drafting.

The missing-context failure is worse than the wrong-fact failure. In the Pakistan case, every individual statement about the gas crisis was true, and the draft was still misleading, because it omitted the surplus that preceded the crisis and the allocation decision that predated it. Facts can all be right while the story is wrong. The control is a deliberate contrarian search before sign-off: look for the strongest evidence against the draft's own framing. In training we put it more bluntly: the model's editorial decisions are more dangerous than its lies, because checking the stated content passes while the missing dimension kills the decision. Two questions catch it: what does this answer assume, and what did it leave out?

5. Honest limits

Five limits, stated plainly because commissioners will ask and should.

This does not replace field verification. The Somalia assessment finished with fifteen named gaps that only key informants can close. A desk product built this way is a structured hypothesis with an interview plan attached, and it should say so on its first page.

Verification costs as much as drafting. Budget it or do not run the workflow. And give it a defined end. The first question every practitioner asks is whether checking is endless. It is not: severity tiers say what must be fixed, a stop condition says when you are done, and confidence is set by what the answer rests on, with the human assigning it.

The analyst owns the judgements. The AI surfaces evidence, flags implications, and attacks drafts. Scenario probabilities, planning ranges, prioritisation, and everything a decision-maker acts on are the analyst's calls, and they should be labelled as such in the output. In training we teach this as a sequence rather than a sentiment: form your own view before the AI reviews the work, check the reviewer's findings against the named sources, and make and communicate the final decision yourself. The reviewer never gets the last word.

One pass is never enough. The audit missed seven dimensions the robustness review found. Assume any single review pass, human or machine, has a comparable miss rate, and structure accordingly. And keep at least one check outside the system that produced the work. A builder and a critic running inside the same box can correct each other out of sight, which is how errors launder themselves into fact. We teach it as a rule: the critic sits inside the same system, so the outside check is still you.

The protocol assumes an analyst who can read it. Writing an attackable hypothesis, judging whether two sources are genuinely independent, and recognising which of the critic's findings are real all draw on domain knowledge the workflow cannot supply. Run by someone without it, the same protocol produces well-formatted confidence instead of verified analysis. The workflow multiplies expertise. It does not replace it. And a note on this paper's own evidence, applying our own scale to ourselves: two cases, one team, implementer self-reported. Grade B. Treat it accordingly, and test it on your own work.

6. What commissioners should require

If your organisation commissions analysis, from consultants, from vendors, or from its own teams, and AI is anywhere in the production chain, the following ten requirements are writable into terms of reference today. They cost the supplier nothing except the ability to sell you the unaudited draft.

1. **Evidence grading on every claim.** Four grades or equivalent, inline, from first draft. An untagged claim is an unaccepted claim.
2. **A stated no-self-citation rule.** Project outputs are never evidence for project outputs.
3. **An adversarial audit log as a deliverable.** Fixed categories, severity tiers, and the fixes applied. The audit is the product's warranty, and a supplier who cannot show one has not run one.
4. **A gap register, not gap concealment.** Unknowns stated, actioned, and handed to field verification with named informant types. Rising gap counts through review are a sign of honesty, not weakness.
5. **A recency sweep in the sign-off week** on every claim about current operational status, with the sweep date recorded.
6. **A contrarian search before sign-off**, documented: what is the strongest evidence against this analysis's framing, and where was it looked for.
7. **Judgement labelling.** Every planning figure, scenario probability, and prioritisation marked as the analyst's judgement, distinguishable from sourced evidence.
8. **A statement of what the document is**, on page one: a desk product produced before field verification is labelled as exactly that.

9. **AI-use disclosure**, stating what the tools did and what the humans did, sufficient for a reader to understand the production chain.
10. **The analyst's name on the judgements.** Accountability does not transfer to a model, and a deliverable whose judgements nobody owns should not be accepted.

None of these requirements slows a good supplier down in any way that matters. Each one removes a place where speed currently hides risk. Together they convert "AI-assisted" from a claim you have to take on trust into a production standard you can inspect. One thing the list cannot do is certify judgement. The requirements make the workflow inspectable. The analyst who reads the outputs is still the thing you are buying.

7. Closing

The question organisations ask about AI-assisted analysis is usually whether the output can be trusted. That question has no general answer. The one that does: did the workflow that produced this output make errors visible, tiered, and fixable before they reached a decision? And that can be answered by inspection, using the requirements above.

The two cases documented here suggest the trade on offer. A week's work in a session, at the price of treating verification as half the job. Speed with the warranty attached. The tools will keep improving, and the failure modes in section 4 will keep occurring, because most of them are not model failures at all. They are the oldest failures in analysis, overstatement, stale facts, missing context, circular sourcing, now arriving at machine speed. The controls are equally old. What is new is that they can now be run as a workflow, every time, without fatigue, by the same systems that create the risk. Or, as we teach it: every AI output is a draft until you decide it is reviewed.

AI-use disclosure

This paper practises Requirement 9. AI tools were used throughout its production: evidence assembly, drafting and revision under my direction, and adversarial audit passes between versions, with a final audit before publication in which the figures were checked against the underlying case records and live sources. The two case studies were run by the MarketImpact team. The protocol design, the analytical judgements and the final text are mine, and I take responsibility for the content. Those audit passes ran as working sessions rather than a preserved log, which is less than Requirement 3 asks of a supplier, and we say so rather than claim otherwise. Anything found after publication will be corrected publicly. MarketImpact's full position on AI use is at marketimpact.org/how-we-use-ai.

AidGPT Guidance Paper MI-GUIDE-002, August 2026. Licensed CC BY 4.0. Prepared ahead of the Markets in Crises Community of Practice fireside "AI on the Ground", 19 August 2026. The workflow this paper documents is taught in the AidGPT programme: aidgpt.org/training. AidGPT Updates, our email list on practical AI adoption, training and governance: aidgpt.org/newsletter.

AidGPT is a training brand of MarketImpact Digital Solutions Ltd. aidgpt.org