AFCEA submission intro (≤70 words):

> *Every year, the leading reason the GAO sustains a bid protest isn't a bad proposal — it's a government evaluation that couldn't be defended against its own criteria. As proposals grow and evaluation teams stay small, consistency and documentation suffer. The answer isn't to remove human judgment from source selection, but to make it more traceable, more consistent, and more auditable.*

Submission image: `image-1-gov.png` (red-pen markup of a strategy document)

The Real Source Selection Risk Isn't the Proposals — It's the Evaluation

By David Kindle, Director of Engineering, Falcon Logic

Ask most people where a federal source selection goes wrong, and they'll point at the proposals: too long, too vague, too aggressive on price. The data tells a different story. Year after year, the leading reason the Government Accountability Office sustains a bid protest is not a flawed offer — it is an unreasonable technical evaluation, followed closely by an unreasonable cost or price evaluation. In other words, the most common point of failure in a source selection is the government's own assessment of what it read.

That pattern has held remarkably steady. In fiscal year 2025, the GAO resolved hundreds of protests on the merits and sustained roughly one in seven, while more than half of all protests filed produced some form of relief once voluntary corrective actions were counted (GAO Bid Protest Annual Report to Congress for Fiscal Year 2025, GAO-26-900695). The specific grounds change little from year to year. They cluster around evaluations that could not be reconciled with the solicitation's own criteria.

Consider a representative example from the GAO's recent findings: an agency credited an awardee with staffing a requirement for the full period when the proposal, read closely, committed to only part of it. The award was reasonable in spirit, perhaps — but the evaluation record did not match the words on the page. That is not a dramatic scandal. It is an ordinary, human lapse of the kind that happens when a small team reads thousands of pages under a deadline. And it is exactly the kind of lapse that turns into a sustained protest, a corrective action, and months of delay for a mission that needed the capability yesterday.

The squeeze that creates the risk

The structural problem is simple. Proposals have grown longer and more complex, spanning multiple volumes across technical, management, past performance, and price. Evaluation teams have not grown to match. The same evaluators are asked to read more, cross-reference more, and reconcile more findings against Section M — all while maintaining perfect consistency between the first proposal they read on Monday and the last one they finish on Friday.

Under that pressure, three things erode. Consistency erodes, because two qualified evaluators can read the same passage and weight it differently. Traceability erodes, because the reasoning behind a rating lives in someone's notes, or in their head, rather than tied to a specific line of the proposal. And documentation erodes, because writing a defensible narrative for every finding is the first task to get compressed when time runs short. Every one of those erosions maps directly onto the grounds that get protests sustained.

Why the obvious fix is the wrong one

It is tempting to reach for general-purpose AI as the remedy — to feed proposals into a model and ask for a summary or a score. That instinct is understandable and, applied carelessly, dangerous. A model that produces a fluent summary with no link back to the source text does not solve the traceability problem; it deepens it. A tool that generates a rating without showing which passage drove it cannot be defended in a protest. And a system that sends sensitive procurement-sensitive material outside a controlled environment introduces a risk far worse than a slow evaluation.

The lesson is not that AI has no place in source selection. It is that the wrong kind of AI makes the core problem worse, while the right kind makes it measurably better.

What responsible AI in source selection should require

If artificial intelligence is going to touch an evaluation record, it should be held to the same standard as a good evaluator — and then some. In practice, that means a few non-negotiable principles.

Every finding must trace to the text. An AI-assisted assessment that cannot cite the specific passage supporting it has no place in the record. Citation is not a feature; it is the whole point.

The human decides; the tool assists. AI can surface, organize, and cross-reference. It should never render the rating. Accountability for the source selection decision has to remain with the people the law makes accountable.

Consistency must be measurable. The value of a tool is that it applies the same reading of the same criteria to every offeror, every time — and can demonstrate that it did.

The data must stay controlled. Procurement-sensitive information belongs inside a secured, government-controlled environment, not in a general-purpose service whose data handling the agency cannot inspect.

The record must be auditable. If the reasoning behind every rating can be reconstructed after the fact, the evaluation is defensible. If it cannot, no amount of speed matters.

The point isn't to judge — it's to make judgment defensible

The promise of AI in acquisition is often sold as speed or headcount savings. That framing misses what actually matters. The enduring problem in source selection is not that evaluations are slow. It is that too many of them cannot withstand scrutiny. The right role for AI is not to replace the evaluator's judgment but to make that judgment more consistent, better documented, and

traceable to the record — so that when an award is challenged, the answer is already written down, finding by finding, in the solicitation's own terms.

Get that right, and the technology does something more valuable than saving time. It protects the integrity of the decision.