



## IBM Defense Model in Practice: What the Model Card Says, and What the Pricing Page Does Not

Janes says the model cuts research time “up to 85%”, with no methodology, baseline, dataset or independent validation attached. That is the loudest number anyone has published about a defense LLM you can run air-gapped.

Here is the confirmed shape: `ibm-defense-3-3-8b-instruct`, roughly 8 billion parameters, built on Granite 3.3 8B, with a 128K-token context length, fine-tuned for defense knowledge and Janes API tool calling, documented for on-premises, air-gapped and classified deployment. IBM announced it on 29 October 2025, Janes on 30 October 2025, and Janes said it was generally available that day. Published benchmark scores: zero. Published prices: zero.



## An 8B Granite derivative with the Janes Inventory API wired in

Strip the announcement language and the product is a Granite 3.3 8B derivative with domain fine-tuning and tool calling pointed at the Janes Inventory API. The fine-tune targets defense-industry knowledge, military doctrine, retrieval-augmented generation, reasoning and instruction following. The tool-calling part is the design decision worth noticing: instead of asking the model to recall platform inventories from frozen weights, it can call Janes for structured data at query time.

That is the correct architecture for this domain. Defense reference data ages fast, and an 8B model's parametric memory is a bad place to store it. Order-of-battle facts and platform variants belong in a queryable source with a provenance trail, not baked into weights someone fine-tuned at an unstated date. IBM and Janes picked retrieval over recall, and that choice does more for trustworthiness than extra parameters would.

The deployment story is the other substantive piece. IBM documents on-premises operation in air-gapped environments, plus classified environments and edge networks. For a buyer whose actual constraint is "nothing leaves the enclave", a model small enough to run on modest hardware with a 128K window for long doctrine documents is a concrete answer.

### **i The version you are actually evaluating**

The documented variant is `ibm-defense-3-3-8b-instruct` on Granite 3.3 8B with 128K context, per the `watsonx docs` model card 2.2.0. IBM's 2.3.x foundation-models overview also lists a separate 3-billion-parameter Defense Model entry, so "IBM Defense Model" is not one fixed size across documentation versions. Pin the exact model ID and doc version in any procurement text you write.

## Why does IBM name benchmarks but publish no scores?

I do not know, and neither does any reader of the published materials. What is confirmed: IBM's materials reference gains on AlpacaEval-2.0 and Arena-Hard plus maths, coding and instruction following. No score. No baseline. No test split. Naming a benchmark and



withholding the number is more conspicuous than saying nothing at all, because it tells you the evaluation was run.

My read: this is probably not concealment of bad results. AlpacaEval-2.0 and Arena-Hard are general instruction-following benchmarks, and a defense fine-tune of an 8B base is unlikely to post embarrassing numbers on them. More likely it is ordinary enterprise caution, plus the fact that those benchmarks say almost nothing about whether the model gets doctrine right or invents a platform variant. The benchmarks that would matter here (domain accuracy against a held-out Janes set, tool-call correctness, refusal behaviour on out-of-scope queries) do not have public standard versions, and building one would mean exposing the corpus.

Which brings up the second gap. Janes does not disclose corpus size, date range, countries or platforms covered, or what share of the fine-tune is attributable to Janes data. For a defense buyer that is not a nice-to-have detail. If your area of responsibility is the South China Sea and the corpus is thin on regional order of battle, the model will still answer fluently. You will not be able to tell from the model card whether it should.

#### THE CLAIM

Janes states the model delivers up to 85% reduction in research time, published on its defence partners page.

#### WHAT IS ATTACHED TO IT

No methodology, no baseline task, no dataset, no sample size, no independent validation. "Up to" sets a ceiling, not an expectation.

I would not treat that figure as hostile marketing. I would treat it as unverifiable, which in a procurement document amounts to the same thing. If someone puts it in a business case, the honest version of the line is: the vendor claims up to 85%, measured by an undisclosed method against an undisclosed baseline.

## A readable model card, an unreadable contract



## IBM Defense Model in Practice: What the Model Card Says, and What the Pricing Page Does Not

### WHAT WORKS

- A published model card you can actually read, with a named base model (Granite 3.3 8B) and a stated context length (128K tokens). Rare in this sector.
- Tool calling to the Janes Inventory API, so structured defense data comes from a live source rather than frozen weights.
- On-premises, air-gapped, classified and edge deployment documented as supported.
- 8B is small enough to be realistic on constrained or disconnected infrastructure.
- Generally available per Janes on 30 October 2025, with no staged rollout gate to wait through.

### ROUGH EDGES

- Zero published benchmark scores, despite AlpacaEval-2.0 and Arena-Hard being named by IBM.
- Zero published prices: no subscription rate, seat fee, token rate or deployment charge in IBM or Janes materials.
- Licensing undisclosed: whether Janes data access is bundled, model-weight redistribution rights, export-control and geographic restrictions all unstated.
- Janes corpus scope undisclosed: size, date range, countries, platforms, and its share of the fine-tune.
- “Air-gapped” and “classified” are described as supported, but no accreditation, authorisation-to-operate status or national classified-information regime is named.
- Two documented parameter sizes (8B and 3B) across doc versions, so the name alone does not identify what you are buying.

The accreditation gap is the one I would push hardest on in a vendor call. “Supports classified environments” is an engineering statement about network topology. It does not mean the model holds an authorisation to operate under any particular national regime, and the materials name none. Those are different questions answered by different people, and conflating them is how a pilot stalls six months in when the accreditation authority asks for paperwork nobody produced.

The licensing silence has a sharper edge than it looks. If Janes data access is bundled with the model, you are buying a data subscription wearing a model’s clothes, and the renewal leverage sits with Janes. If it is separate, the tool calling that makes the whole architecture sensible stays inert until you sign the second contract. Nobody has published which it is. On the weights side, redistribution rights matter for anyone planning to push the model to



edge devices a partner nation operates, and that question is unanswered too. Same pattern I wrote about in [open weights versus open source](#): the artifact looks inspectable, the terms governing it are not.

#### MY TAKE

*The twelve-day gap between the product page going live on 17 October 2025 and the newsroom announcement on 29 October 2025 is a small tell. Product pages usually ship ahead of press when sales conversations are already running. Unconfirmed, but my working assumption is that this was sold to named accounts before it was announced, which would explain the absent price list: pricing here is negotiated per engagement, not published. Normal for defense procurement. It also means a public evaluator has no anchor at all.*

## Pilot it if you already own Janes, skip it if you are buying on published evidence

If you already hold a Janes subscription and your analysts spend their days inside that data, this is worth a scoping call with IBM. The tool-calling integration is the real product, and you are the buyer it was built for. Ask for the benchmark numbers under NDA, ask for corpus scope by region, ask whether data access is bundled.

If you need a defense-capable model in an air-gapped enclave and your alternative is self-hosting a general open-weights model plus your own RAG layer, run both. The honest comparison is whether IBM's fine-tune plus the Janes connector beats your own retrieval stack over the same corpus, and published material will not answer that. Budget for a bake-off.

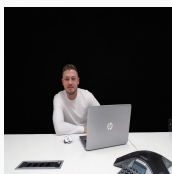
Skip it if you are evaluating on published evidence alone. There is no score to compare, no price to model, no corpus description to match against your mission area. The documentation is better than most military AI products offer, which is a low bar, and it still does not contain a single figure that would survive a procurement review. \*\*The gap between documented and evaluable is the whole story here\*\*, the same way [a prerelease flag tells you more than a docs page](#).

One practical note before you write anything binding: the 3B entry sits in the 2.3.x



## IBM Defense Model in Practice: What the Model Card Says, and What the Pricing Page Does Not

foundation-models overview while the 8B card lives under 2.2.0, so re-read both and record which doc version you read. The model name on the invoice and the model card you evaluated should match, and right now that is not automatic. If you are taking this to a procurement board and want a second pair of eyes on which undisclosed items are genuinely blocking and which you can negotiate around, [send me the shortlist and we can talk it through](#).



WRITTEN BY

**Artur Markus**

Software engineer in Basel building AI infrastructure and agentic systems, with a background in pharmaceutical automation, GxP compliance, and validation. I help teams turn AI from a chatbot into a reliable operational layer.

**Book a meeting →**