

RAG readiness checklist

Fifteen checks we run before agreeing to build retrieval over a document corpus. Tick what you already have and read the result on page two.

The corpus: what you actually have

- ☐ **01. You can name the exact place where the documents live.**
"Somewhere on the drive and in Marina's inbox" is not a corpus.
- ☐ **02. You know how many there are: hundreds, thousands or tens of thousands.**
The order of magnitude decides chunking and retrieval strategy.
- ☐ **03. You know what share are scans and photos rather than text.**
A scan is a separate recognition stage with its own cost and error rate.
- ☐ **04. You can say which version of a document is current.**
The system quotes a superseded revision as confidently as the live one.

Structure: how a document is built

- ☐ **05. Documents have section headings rather than one block of text.**
Headings are the cheapest boundary to chunk on.
- ☐ **06. You know whether there are tables and how much answers depend on them.**
A table cut in half produces a confidently wrong answer.
- ☐ **07. You know whether documents cross-reference each other.**
"As set out in Appendix 3" is useless without Appendix 3.
- ☐ **08. Duplicates and near-duplicates are known at least approximately.**
Five versions of one policy crowd everything else out of the results.

Permissions: who sees what

- ☐ **09. You know whether every document may be shown to every employee.**
Permissions must be designed into retrieval from day one.
- ☐ **10. You know whether the corpus contains personal data.**
It decides where the corpus may be processed and which model is allowed.
- ☐ **11. There is a person accountable for the content of the documents.**
With no content owner a dispute about a wrong answer hits the developer.
- ☐ **12. You know how often the documents change.**
That decides whether incremental re-indexing is needed.

Verifiability: how you will know it worked

- ☐ **13. You can write down twenty real questions people actually ask.**
Not invented ones. This is the actual specification.
- ☐ **14. For each question you know which document holds the answer.**
It turns "I like it or I don't" into something measurable.
- ☐ **15. There is an agreement on what counts as an acceptable answer.**
Right document in the top three, a source link, an honest "not found".

How to read your result

12 or more yeses

The data is ready. Start with a pilot on one section of the corpus and your twenty questions: two or three weeks show whether retrieval lands.

7 to 11

The normal state of a working company. Start with the part of the corpus with the most yeses while closing permissions and versioning gaps.

Fewer than 7

RAG right now gives a beautiful demo and a system that does not work. Spend two or three weeks putting documents in order first: owner, versions, current revisions. That work is not about AI.

Done the checklist?

Tell us how it went: how many documents, which answers were no, and what questions the system must close. You get a review of your corpus in return — what to do first and whether you need RAG at all.

ioann-ai.ru/en/tools/rag-readiness-checklist