5 Questions your AI vendor hopes you won't ask

From Architecture of Intellect by Igor Ageyev • **notesofea.com**

1. **Where is the routing layer?**

   If every request, simple or complex, goes to the same large model, you're paying premium rates for commodity work. The architecture was designed for the vendor's revenue, not your economics.

2. **What is the escalation boundary, mathematically?**

   "A human reviews low-confidence outputs" is a sentence. A defined confidence threshold with measured false-accept rates is an engineering answer. Which one do they have?

3. **Are the prompts version-controlled and regression-tested?**

   A system prompt is production logic. If it's edited like a memo rather than deployed like code, your system's behaviour changes with no audit trail.

4. **Can you trace any output back to its sources?**

   Prompt version, retrieved documents, model version. High-risk systems under the EU AI Act must produce exactly this lineage. "The computer said so" is not a defence; it's an admission.

5. **What halts deployment?**

   If a model update fails their internal evaluation, does the release stop automatically? A vendor with no answer has no evaluation, only a changelog.

How to read the answers:

You don't need to understand them technically. Watch the comfort. Engineers who have answers enjoy these questions.

The full architect's audit and the architecture behind it is in Architecture of Intellect:
www.notesofea.com/book