GenAI Insights

01 / 11

Prompt engineering isn't dead.

It just got demoted.

Why context engineering is the real job in 2026.

das-purnendu

GenAI Insights

02 / 11

The Evolution of AI Skills

From prompts to orchestration.
Here's how we got here →

2023 — **Prompt engineering**
Clever wording.
Better outputs.

2024 — **RAG pipelines**
Give the model
the right facts.

2025 — **Tool use & agents**
Give the model
actions to take.

2026 — **Context engineering**
Orchestrate the
whole stack.

das-purnendu

GenAI Insights

03 / 11

What is Context Engineering?

> The delicate art and science of filling the context window with just the **right information** for the **next step.**
>
> — **Andrej Karpathy,** June 2025

The Context Window

Tobi Lütke (Shopify CEO) and Simon Willison shaped early usage. The term stuck because "**prompt engineering**" had been diluted to mean "**typing into ChatGPT.**"

GenAI Insights

04 / 11

Why This Skill Matters Now

Three forces are converging.
Context engineering is the leverage.

1. AGENTS NEED CONTEXT

Agents need state, memory, tool definitions, and history — all in context.

2. TOKENS = COST + LATENCY

Context windows are finite, and tokens scale linearly with cost (worse with latency).

3. ENTERPRISE REALITY

BCG: ~70% of enterprise AI failures trace to missing context, not weak models.

~70% of enterprise AI failures are due to missing context, not weak models.

GenAI Insights

What Lives in Your Context Window?

The model sees only what you put inside.
Every slot **matters**.

1. System Instruction
(the how)

2. Task / User Query
(the what)

3. Few-shot Examples
(the like this)

4. Retrieved Knowledge
(RAG, schemas, docs)

5. Memory
(short-term, long-term, episodic)

6. Tool Definitions
+ Prior Tool Outputs

Context Window

Every slot is a knob.
Context engineering = tuning all six.

GenAI Insights

06 / 11

Pillar 1: Selection

What you pull in.

BAD

Dump every document that matched the keyword.

More context ≠ better results.
It adds noise and cost.

GOOD

Dense retrieval → reranker → top-K → trust filter.

RULE: Relevance beats volume, always.

Most “RAG is broken” complaints are actually **retrieval-tuning** complaints.

Focus on:

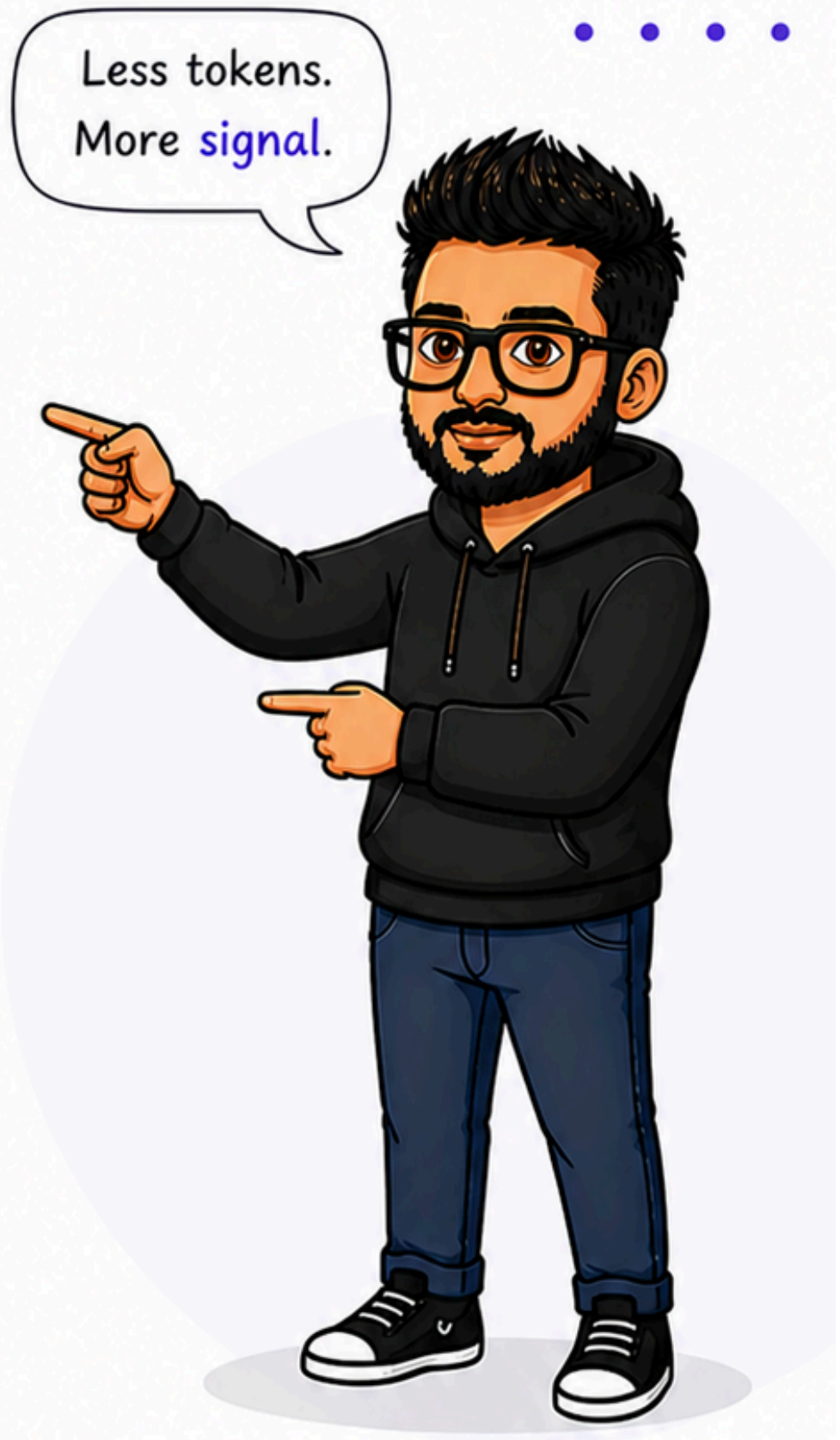
- Better queries
- Better retrieval
- Better filtering

GenAI Insights

07 / 11

Pillar 2: Compression

Token budget is real money.



1. Hierarchical Summarization

Turn long docs into layered summaries. Keep the important, drop the fluff.

2. Schema Compression

Column pruning, type abbreviations, schema-as-DSL. Give the model only what it needs.

3. Prompt Compression

Use prompt-compression libraries (LLMLingua and friends) to shrink prose without losing meaning.

4. Prefer Structured > Prose

Tables, lists, schemas, JSON > paragraphs, wherever possible.

Every token you save buys you:

- Lower cost
- Faster responses
- More room for what actually matters

RULE: If a list works, never paragraph it.

BAD (Prose)

GOOD (List)

GenAI Insights

Measure What Matters (RAG)

You can't improve what you don't measure.
Track these **key metrics**.

1. Answer Relevance

% of responses rated relevant by users.

85%

2. Retrieval Quality

Precision@K, Recall@K, or MRR of retrieved docs.

3. Hallucination Rate

% of factual errors or unsupported statements.

Low

4. Latency (P95)

End-to-end response time at the 95th percentile.

5. Cost per 1K Queries

Track LLM, embedding, and infra costs per 1K user queries.

Build a simple dashboard.
Data-driven RAG > Guess-driven RAG.

GenAI Insights

09 / 11

BONUS:
3 Advanced Tips to Level Up Your RAG

Go from good RAG to **production-ready** RAG.

1. Hybrid Search (Best of Both)

- Combine keyword + vector search
- Improves recall on names, numbers, and facts
- Don't rely on vector search alone

Keyword Search + Vector Search → Better Results

2. Rerank with a Strong Model

- Use cross-encoder rerankers
- Reorder top 50–100 docs
- Quality boost with minimal cost

Top 50 Docs → Reranked Top 5

3. Evaluate Continuously

- Build a test set of real queries
- Track metrics (slide 8)
- Iterate, measure, improve

Evaluate → Improve → Measure

Great RAG is not luck.
It's engineered.

- Better data
- Better retrieval
- Better results

GenAI Insights

10 / 11

Key Takeaways: Build RAG That Actually Works

Great RAG = Right content, right context, right evaluation.

1 Focus on Quality Over Quantity

- Right data > More data
- Filter, chunk & rank smarter
- Precision always beats recall

2 Context is Your Superpower

- Better context = Better answers
- Keep it fresh, relevant & scoped
- Right context, right audience

3 Measure, Iterate & Level Up

- Track what actually matters
- Iterate based on data, not guess
- Small improvements compound fast

Build. Measure. Improve.
That's how you build RAG that wins.

Your next version will be **10x better** than the last.

das-purnendu

GenAI Insights

That's a Wrap!

Build. Measure. Improve.

Great RAG isn't magic.
It's a system you **build**, **test** and **refine**.

Loved this content?
Let's grow together!

Like
If you found this useful

Comment
Share your thoughts or questions

Share
Help others who want to level up

Save
Save this post for future reference

Follow @das-purnendu
for more practical GenAI insights and **real-world** content!

Follow