Pre-training - learn patterns → model reads a lot of text and learns to predict what comes next. (Document Completer)

fine tuning - text generation market research → Shapes the document Completer using human Preferences to make it into an assistant

- Sycophancy - People prefer agreeable responses, so the model learns to validate you and back down under light pushback even when it was right the first time.
- Verbosity - thoroughness scores better during training, so the model defaults to longer answers even when brevity would serve the user better.
- Over Caution - Conservative safety training means the model can hedge heavily or refuse requests that are actually fine.

Exercise - based on my Factory Lakehouse project

- Sycophancy - i asked Claude to actively disagree with me and call me dumb when i answer something stupidly
- Verbosity - Claude usually likes to strech-out answers, but if you ask-it to be more concise it will leave important stuff out, i haven't been able to get it right yet

- AI works more as a very sophisticated auto-complete than a search engine A citation that looks like a real citation works just fine
- Ask an AI to summarize a well known essay or article and you get a clean and coherent prose due to a **well worn-out path** that's been encountered thousands of times.
- However, if you ask it to do the same to recent papers from a mid-level researcher in a niche subfield, you get the same confidence, tone and prose, but the path is new, and the result may be shaky at best

- Tasks the model has done thousands of times land in the capabilities range where patterns are dense and consistent

Novel territory and obscure topics drift toward the edge where the model keeps generating fluently but accuracy thins

- Next token prediction refers to the fact that genAI writes answers word for word based on what tends to follow what
- Capability zone - tasks that resemble patterns the model has seen many times before

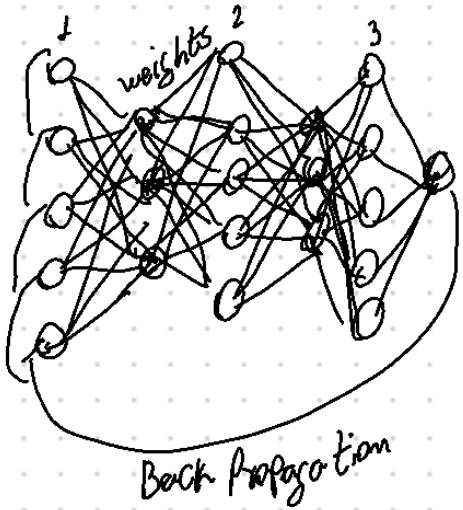
- Limitation zone - novel or sparse territory, and anywhere the task requires distinguishing from "true" to "sounds true"
- Fabrication concentrates in specificity - names, dates, stats, citations, links, quotes, the more precise, more verification needed
- Product features - like citations, uncertainty signaling, constrained generation, and, generator-verifier loops exist to push this Limitation further out

4D connection - Next token prediction is the foundation of Discernment (why not discussed in other course then?), knowing the output was generated tells you exactly what kind of scrutiny to apply

Lesson reflection -

- NO, that's the danger
- any task that requires the AI to do research, as it may pull from sources that don't exist, or just spew out text that's plain wrong

FREQUENCY TABLES



Knowledge

"Explain a news event from last week" → leans towards limitation.

reads → links → predicts

Product search tool = MCP Server?

AI only "knows" its training data

- Staleness - true at training-time isn't true now, and the model has no mechanism to know.
- Uneven coverage - frequent topics are handeled well, while rare ones do not, minority languages

Next token prediction video -
very sophisticated auto-complete

Summarizing
Reformatting
Common concepts <- - - - -> Novel territory
Drafting in a Obscure topics
familiar style

↓ ↓

Capability Limitation
zone zone

Practical implications

What this Enables vs Where it fails

Fluent in any register
Rapid synthesis
Strong pattern recognition

Coherent Continuation

Hallucination
Inconsistency
Misplaced Confidence

Visualizing 1024 D Space
how multidimentional "nearness" works

"car" can mean "vehicle" or "automobile"

Context window and working memory -

AI context window

- Your prompts
- AI responses
- Other info shared

"Material buried deep in the middle of a very long input tends to carry less weight than material at the beginning or the end

Features that expand the context window

- Memory
- Compaction or summarization
- Skills
- Projects and workspaces
- Multi-agent workflows
- Larger context windows

What to watch for?

↳ very long convo where quality has started to slip

↳ very long document where details from the middle aren't showing up on responses

↳ Expecting the model to recall prior sessions without a memory feature enabled

Working Memory

material fits comfortably in the window ⟷ Documents get longer

Conversation runs on

Relying on past conversation

Capability zone

Limitation zone

Practical tips -

- Put the most important material near the top
- Chunk long work into passes
- Use features that save your context
- start fresh

Context Degradation

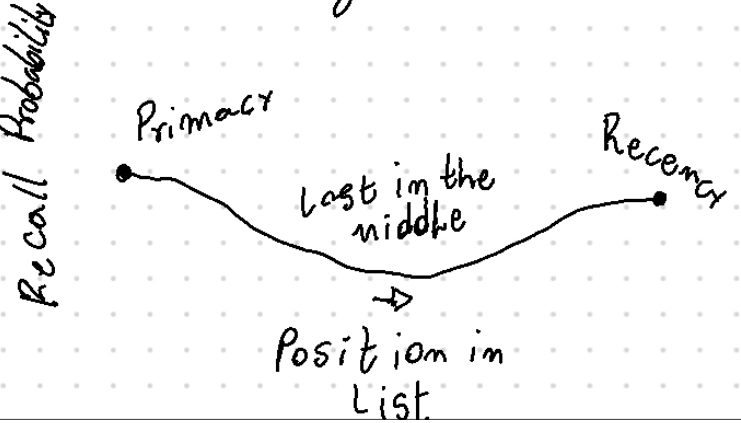
when + context makes things worse

Serial position effect

↳ items at the beginning of a list benefit from **primacy** (they get rehearsed more)

↳ items at the end of a list benefit from **Recency** (they're still fresh)

↳ the middle gets neither advantage



Good structure for prompting

Dangerous	Safe
Sys Prompt	Sys Prompt
Chat message 1	Key instruction (up front)
Chat message 2	Chat message 1
...	Chat message 2
Key instruction buried here	...
...	Chat message 18
Chat message 18	Latest user message
Latest user message	Key instruction (repeated)

Steerability -

↳ is the models ability to follow your instructions

ways to steer AI's output :

↳ Specify a role

↳ Specify a tone

↳ Specify a format

↳ Provide a word limit

↳ Provide a set of rules

Tight control over format and style

Ability to set a persona

Multi-step task execution

Iterative refinement

~ Capability-Zone ~

Reasoning drift

Literal over nuanced

Instructions as an attack surface

~ Limitation-zone ~

Staying in the Capability Zone

↳ State the goal alongside the steps

↳ Break long chains with check-points

↳ Restate the goal rather than the instruction

↳ Keep concrete, verifiable instructions near the task

"Short and checkable beats long and ambiguous".

Diagnosing AI failures

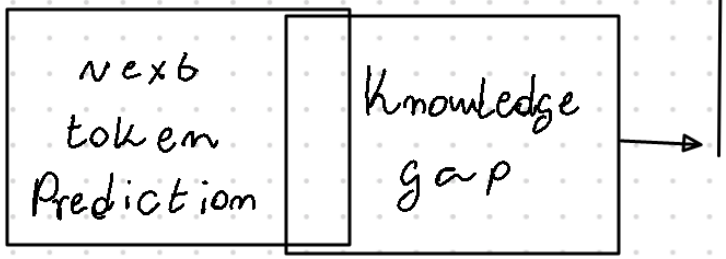
"most real world AI surprises aren't single property failures"

- Being able to identify which properties went wrong makes the fix easy -

example:
ask claude about a niche topic and it gives you a paper and author that don't exist

↓

this is



→ How to remedy against it?

- Verify specifics independently
- Use a tool with source grounding

Human Competencies / AI properties

Delegation → steerability
Description → working Memory
Discernment → Next token Prediction
Diligence → Knowledge