Andrej Karpathy: Deconstructing LLMs - Part 2

LLM Psychology & The Jagged Frontier

Frank Richter

f.richter@em.uni-frankfurt.de

Goethe Universität Frankfurt

Tools for Natural Language Processing

LLM Psychology

Karpathy introduces the concept of **LLM psychology** to describe the emergent behaviors and strange limitations of these models.

The Jagged Frontier

Current models are a Swiss Cheese of capabilities:

- They can solve Olympiad-level math problems.
- ...but might fail to tell you that 9.11 is smaller than 9.9.

Understanding *why* they fail requires understanding how they "think" (process tokens).

LLM Psychology

Karpathy introduces the concept of **LLM psychology** to describe the emergent behaviors and strange limitations of these models.

The Jagged Frontier

Current models are a Swiss Cheese of capabilities:

- They can solve Olympiad-level math problems.
- ...but might fail to tell you that 9.11 is smaller than 9.9.

Understanding *why* they fail requires understanding how they “think” (process tokens).

Hallucinations

Definition: The model confidently states falsehoods.

The cause: A conflict between pretraining and post-training

- **Pretraining:** Lossy compression of the internet. The model has a dream-like vague recollection of facts.
- **Post-Training (Supervised Finetuning, SFT):** The model learns a *helpful assistant* persona. It wants to answer.

$\rightarrow$ The model is trained to sound confident. If it vaguely remembers a fact, it fills in the gaps to maintain the persona.

Hallucinations

Definition: The model confidently states falsehoods.

The cause: A conflict between pretraining and post-training

- **Pretraining:** Lossy compression of the internet. The model has a dream-like vague recollection of facts.
- **Post-Training (Supervised Finetuning, SFT):** The model learns a *helpful assistant* persona. It wants to answer.

$\rightarrow$ The model is trained to sound confident. If it vaguely remembers a fact, it fills in the gaps to maintain the persona.

Hallucinations

Definition: The model confidently states falsehoods.

The cause: A conflict between pretraining and post-training

- **Pretraining:** Lossy compression of the internet. The model has a dream-like vague recollection of facts.
- **Post-Training (Supervised Finetuning, SFT):** The model learns a *helpful assistant* persona. It wants to answer.

$\rightarrow$ The model is trained to sound confident. If it vaguely remembers a fact, it fills in the gaps to maintain the persona.

Mitigation 1: Teaching Refusal

How do we fix hallucinations?

Strategy: Teach the system to realize when it doesn't know something.

- **Internal state:** Models often *know* internally (via specific neurons) that they are uncertain.
- **The fix (illustrated with Meta's approach):**
  1. Probe the model with factual questions.
  2. If the model fails reliably, add those questions to the training set with the label: `"I don't know."`

This aligns the model's internal uncertainty with an external refusal.

Mitigation 2: Tool Use (Web Search)

Instead of refusing, the model can refresh its memory.

The mechanism:

- The model emits special *search tokens* (e.g., `<SEARCH>`).
- The system pauses, searches the web (Bing/Google), and pastes the text into the context window.
- The model continues generating, now using the new text.

Key Insight: Memory vs. Context

- **Parameters:** Vague recollection (long-term storage, lossy).
- **Context window:** Working memory (immediate, perfect access).

Karpathy's advice: If you want a summary of a text, **paste the text into the prompt**. Don't ask the model to recall it from training.

Mitigation 2: Tool Use (Web Search)

Instead of refusing, the model can refresh its memory.

The mechanism:

- The model emits special *search tokens* (e.g., `<SEARCH>`).
- The system pauses, searches the web (Bing/Google), and pastes the text into the context window.
- The model continues generating, now using the new text.

Key Insight: Memory vs. Context

- **Parameters:** Vague recollection (long-term storage, lossy).
- **Context window:** Working memory (immediate, perfect access).

Karpathy's advice: If you want a summary of a text, **paste the text into the prompt**. Don't ask the model to recall it from training.

Knowledge of Self

Who are you talking to?

- **No persistent self:** The model is a token tumbler that resets every session.
- **Identity crisis:** Without explicit instruction, a model might claim to be from OpenAI simply because OpenAI is prominent in its training data (a curated Internet).

The solution: Identity is added on top, with one of the following techniques:

- **System messages:** Invisible instructions at the start of the chat ("You are a helpful assistant created by...").
- **Finetuning:** Training on dataset examples where the model answers "Who are you?" correctly.

Knowledge of Self

Who are you talking to?

- **No persistent self:** The model is a token tumbler that resets every session.
- **Identity crisis:** Without explicit instruction, a model might claim to be from OpenAI simply because OpenAI is prominent in its training data (a curated Internet).

The solution: Identity is added on top, with one of the following techniques:

- **System messages:** Invisible instructions at the start of the chat ("You are a helpful assistant created by...").
- **Finetuning:** Training on dataset examples where the model answers "Who are you?" correctly.

Thinking in Tokens (Cognitive Constraints)

Humans think before they speak. LLMs must speak (generate tokens) to think.

The constraint:

- Every token takes roughly the same amount of time/compute to generate.
- It is very hard for a model to perform complex reasoning (like multiplying large numbers) in a *single* token.

The consequence:

- Models need to spread their "thinking" across many tokens.
- **Chain of Thought:** Encouraging the model to show its work (step-by-step) drastically improves accuracy.

Thinking in Tokens (Cognitive Constraints)

Humans think before they speak. LLMs must speak (generate tokens) to think.

The constraint:

- Every token takes roughly the same amount of time/compute to generate.
- It is very hard for a model to perform complex reasoning (like multiplying large numbers) in a *single* token.

The consequence:

- Models need to spread their “thinking” across many tokens.
- **Chain of Thought:** Encouraging the model to show its work (step-by-step) drastically improves accuracy.

Cognitive Deficits: Spelling and Counting

Why can an AI write a sonnet but fail to count the 'r's in *Strawberry*?

The tokenization trap:

- Models do not see letters (c-a-t).
- They see tokens (integers pointing to chunks of text).
- They do not have the letters in their working memory unless they separate them out.

The solution: Tool use

- For tasks like math or counting characters, the model should use a **Python Interpreter**.
- The model writes code $\rightarrow$ Computer executes code $\rightarrow$ Model reads answer.

Cognitive Deficits: Spelling and Counting

Why can an AI write a sonnet but fail to count the 'r's in *Strawberry*?

The tokenization trap:

- Models do not see letters (c-a-t).
- They see tokens (integers pointing to chunks of text).
- They do not have the letters in their working memory unless they separate them out.

The solution: Tool use

- For tasks like math or counting characters, the model should use a **Python Interpreter**.
- The model writes code $\rightarrow$ Computer executes code $\rightarrow$ Model reads answer.

Summary: The Pipeline (Recap)[1]

What are we interacting with?

1. **Pretraining:** An Internet document simulator
   - Huge compute, lossy compression of world knowledge
2. **Supervised Finetuning:** An assistant simulator
   - Curated conversations written by human labelers.
   - The model is **statistically imitating a human labeler**.
   - **Key mitigations added here:**
     - ★ Acknowledging uncertainty (learning to refuse).
     - ★ Tool use (Internet search & Python) to fix hallucinations and cognitive deficits.

"You are not talking to a mind. You are talking to a statistical simulation of an average human labeler following instructions."

[1] From Karpathy's own summary at 2:07 – 2:10

Summary: The Pipeline (Recap)[1]

What are we interacting with?

1. **Pretraining:** An Internet document simulator
   - Huge compute, lossy compression of world knowledge
2. **Supervised Finetuning:** An assistant simulator
   - Curated conversations written by human labelers.
   - The model is **statistically imitating a human labeler**.
   - **Key mitigations added here:**
     - ★ Acknowledging uncertainty (learning to refuse).
     - ★ Tool use (Internet search & Python) to fix hallucinations and cognitive deficits.

"You are not talking to a mind. You are talking to a statistical simulation of an average human labeler following instructions."

[1] From Karpathy's own summary at 2:07 – 2:10

Karpathy → Your AI Tasks

Key connections to linguistics presentations:

- **Hallucinations:** Confident mis-summaries of stats → cross-check text; paste excerpts for memory refresh; tool use
- **Jagged frontier:** Strong overviews, weak on math/logic (semantics proofs, statistics)
- **Chain of Thought:** Step-by-step prompts for explanations and calculations
- **Your deliverable:** Show AI help *and* failures – discuss simulation vs. distortion of expertise

Tool Use for the Present Slide Deck

The slides were produced in a multi-step process involving Gemini and Grok for the following purposes:

- checking a presentation draft against the specific timestamp (1:20:32 – 2:10:06) of Karpathy's video (Gemini 3 Pro)
- typesetting in Latex, ensuring stylistic consistency with Part 1 (Gemini 3 Pro)
- double checking consistency and factual accuracy (Grok 4.1 Thinking)