Andrej Karpathy: Deconstructing LLMs - Part 3

Reinforcement Learning & Future Directions

Frank Richter

f.richter@em.uni-frankfurt.de

Goethe Universität Frankfurt

Tools for Natural Language Processing

From the Previous Recap

What are we interacting with?

1. **Pretraining:** An Internet document simulator
   - Huge compute, lossy compression of world knowledge
2. **Supervised Finetuning, SFT:** 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

Karpathy's intermediate conclusion:

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

Reconceptualization: Taking an LLM to School

Karpathy describes **three** training stages using the metaphor of a student learning from a textbook.

1. **Pretraining → the reader**
   - *Analogy:* Reading the exposition chapters of the textbook
   - *Result:* Broad knowledge acquisition → base model
2. **Post-Training 1, SFT → the imitator**
   - *Analogy:* Studying the **worked examples**
   - *Result:* Learning how to answer like an expert → assistant model
3. **Post-Training 2, RL → the thinker**
   - *Analogy:* Doing the **practice problems**
   - *Method:* Trial and error; checking against the answer key
   - *Result:* Learning *how* to think → reasoning model

Reconceptualization: Taking an LLM to School

Karpathy describes **three** training stages using the metaphor of a student learning from a textbook.

1. **Pretraining → the reader**
   - *Analogy:* Reading the exposition chapters of the textbook
   - *Result:* Broad knowledge acquisition → base model
2. **Post-Training 1, SFT → the imitator**
   - *Analogy:* Studying the **worked examples**
   - *Result:* Learning how to answer like an expert → assistant model
3. **Post-Training 2, RL → the thinker**
   - *Analogy:* Doing the **practice problems**
   - *Method:* Trial and error; checking against the answer key
   - *Result:* Learning *how* to think → reasoning model

Stage 3: Reinforcement Learning (RL)

Why isn't SFT enough?

- SFT relies on *imitation*.
- Humans (labelers) do not know the optimal way for an AI to "think" (process tokens) to solve a hard problem.

The RL approach in verifiable domains:

- **Domain:** Math, code, games (such as Go)
- **Method:** Guess and check
  1. The model generates many solutions (rollouts).
  2. Solutions are checked against a ground truth (the correct answer).
  3. Successful paths are reinforced.

Stage 3: Reinforcement Learning (RL)

Why isn't SFT enough?

- SFT relies on *imitation*.
- Humans (labelers) do not know the optimal way for an AI to "think" (process tokens) to solve a hard problem.

The RL approach in verifiable domains:

- **Domain:** Math, code, games (such as Go)
- **Method:** Guess and check
  1. The model generates many solutions (rollouts).
  2. Solutions are checked against a ground truth (the correct answer).
  3. Successful paths are reinforced.

Thinking Models: DeepSeek R1 et al

When models do *true RL* on difficult problems, we observe emergent properties.

Chain of Thought, CoT:

- The model learns to produce long sequences of *thinking tokens* before the final answer.
- **Behavior:** Self-correction, backtracking, re-evaluating ("Wait, that's not right...").
- **Discovery:** No human taught it to "think" this way. The optimization process discovered that more thinking tokens lead to higher accuracy.

Speculation on RL: The AlphaGo Moment

Can AI exceed human intelligence?

SFT ceiling: If you only train on human data (imitation), you can never surpass the human expert like the Go champion Lee Sedol.

An RL breakthrough (Move 37):

- By playing against itself and optimizing for the *win* (reward), AlphaGo discovered strategies humans had never seen.
- **Implication:** In verifiable domains (Math/Code), RL allows models to discover knowledge and strategies beyond human capability.
- **Speculation:** Could lead to superhuman strategies in code/math, but unproven in language yet.

Speculation on RL: The AlphaGo Moment

Can AI exceed human intelligence?

SFT ceiling: If you only train on human data (imitation), you can never surpass the human expert like the Go champion Lee Sedol.

An RL breakthrough (Move 37):

- By playing against itself and optimizing for the *win* (reward), AlphaGo discovered strategies humans had never seen.
- **Implication:** In verifiable domains (Math/Code), RL allows models to discover knowledge and strategies beyond human capability.
- **Speculation:** Could lead to superhuman strategies in code/math, but unproven in language yet.

Unverifiable Domains, Creative Writing

What if there is no right answer? ("Write a funny joke about pelicans.")

Reinforcement Learning from Human Feedback, RLHF:

1. Humans rank model outputs (joke A > joke B).
2. Train a **reward model** to simulate the human judge.
3. Run RL against this reward model.

The limit (Goodhart's law):

- *Concept:* "When a measure becomes a target, it ceases to be a good measure." (Marilyn Strathern, Charles Goodhart)
- If you optimize too hard against the reward model, the trained model games it and produces nonsense like "The the the..." that the reward model mistakenly loves.

RLHF is fine-tuning, not open-ended improvement.

Unverifiable Domains, Creative Writing

What if there is no right answer? (“Write a funny joke about pelicans.”)

Reinforcement Learning from Human Feedback, RLHF:

1. Humans rank model outputs (joke A > joke B).
2. Train a **reward model** to simulate the human judge.
3. Run RL against this reward model.

The limit (Goodhart’s law):

- *Concept:* “When a measure becomes a target, it ceases to be a good measure.” (Marilyn Strathern, Charles Goodhart)
- If you optimize too hard against the reward model, the trained model games it and produces nonsense like “The the the...” that the reward model mistakenly loves.

RLHF is fine-tuning, not open-ended improvement.

Unverifiable Domains, Creative Writing

What if there is no right answer? ("Write a funny joke about pelicans.")

Reinforcement Learning from Human Feedback, RLHF:

1. Humans rank model outputs (joke A > joke B).
2. Train a **reward model** to simulate the human judge.
3. Run RL against this reward model.

The limit (Goodhart's law):

- *Concept:* "When a measure becomes a target, it ceases to be a good measure." (Marilyn Strathern, Charles Goodhart)
- If you optimize too hard against the reward model, the trained model games it and produces nonsense like "The the the..." that the reward model mistakenly loves.

RLHF is fine-tuning, not open-ended improvement.

Future Directions

Where is the field going?

- **Multimodality ("Omni-models"):**
  - Current: Stitching models together (Speech-to-Text $\rightarrow$ LLM $\rightarrow$ Text-to-Speech).
  - Future: **Native** multimodality. Audio and images are just tokens. The model hears tone and sees pixels directly.
- **Agents:** Moving from chat to jobs. Models that can use computers, browse the web, and execute long-horizon tasks over hours.
- **Test-time training:** Updating the model's *brain* temporarily during inference (learning while working), rather than just relying on frozen parameters.

Summary: The Karpathy Stack

1. **Pretraining:** The Internet simulator, knowledge
2. **SFT:** The assistant simulator, personality & format
3. **RL (verifiable):** The thinker, reasoning & improved accuracy
4. **RLHF (unverifiable):** The preference optimizer, human preference alignment

Conclusion: After RL the models are no longer pure simulations of human labelers; they develop their own "thinking" patterns.

The Jagged Frontier Remains

Even with RL, the models of course remain inherently stochastic. They are tools, and they can fail catastrophically. Always verify the output.

Summary: The Karpathy Stack

1. **Pretraining:** The Internet simulator, knowledge
2. **SFT:** The assistant simulator, personality & format
3. **RL (verifiable):** The thinker, reasoning & improved accuracy
4. **RLHF (unverifiable):** The preference optimizer, human preference alignment

Conclusion: After RL the models are no longer pure simulations of human labelers; they develop their own “thinking” patterns.

The Jagged Frontier Remains

Even with RL, the models of course remain inherently stochastic. They are tools, and they can fail catastrophically. Always verify the output.

Karpathy → Your Presentations

Applying RL insights to your AI tasks:

- **Reasoning models:** Use CoT prompts for paper breakdowns (semantics proofs, stats methods); never forget to verify outputs.
- **Verifiability:** In "unverifiable" domains like linguistics, watch for Goodhart's gaming (like over-optimized summaries distorting arguments).
- **Jagged frontier:** Demo RL-like self-correction in AI, plus failures; discuss superhuman potential vs. human oversight.
- **Your goal:** Show AI as thinker/simulator for papers, but always cross-check like RL "ground truth."

Tool Use for the Present Slide Deck

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

- generating a stylistically consistent presentation draft in LaTeX covering the specific video segment (2:07:28–end) (Gemini 3 Pro)
- performing several rounds of revisions and cross-checking (Gemini 3 Pro)
- double-checking consistency and accuracy; calibrating content against the course syllabus; and considering further ideas (Grok 4 Expert)