Andrej Karpathy: How I use LLMs?

Key Objectives

- Get a solid grasp of how LLMs work—from breaking down text into tokens and managing context windows to understanding the various training stages.
- Dive into the diverse ecosystem of LLM providers and learn about the different pricing models they offer.
- Find out how to make the most of integrated tools—like web search, code execution, and file uploads—to boost the quality of outputs.
- Explore the exciting world of multimodal interactions (voice, image, and video) along with customization options.

---

1. Fundamentals of LLM Operations

1.1 Tokenization & Context Windows

- **Definition:**
  - **Tokenization:** This is the process of splitting text into small units known as **tokens**.
  - **Context Window:** Think of it as a one-dimensional sequence of tokens that serves as the model's temporary memory.
- **Key Points:**
  - Every query you send and every response the model gives is broken down into tokens.
- **Practical Tip:**
  - Keep the context window light. Overloading it with too many tokens can distract the model and slow things down. When changing topics, starting a new chat helps reset the context.

1.2 Pre-Training vs. Post-Training

- **Pre-Training:**
  - During this phase, the model absorbs huge amounts of internet text, compressing it into a "lossy, probabilistic zip file" of neural network parameters.
  - **Outcome:**
    * This results in a fixed knowledge cutoff—meaning the model only knows what was available up to its last training session.
- **Post-Training:**
  - Here, the model is fine-tuned with human-curated conversation datasets to give it personality and a more natural, conversational tone.
  - **Result:**
    * The model evolves into a self-contained entity that mixes its stored (pre-trained) knowledge with a humanlike style (post-training).
- **Analogy:**
  - Imagine the model as a "zip file" that not only stores a massive amount of data but also carries a personality shaped by real human examples.

---

2. The LLM Ecosystem

2.1 Providers and Model Variants

- **Major Players:**
  - **OpenAI's ChatGPT:** The leader in the field and widely recognized, often dubbed the "Original Gangster" of LLMs.
  - **Other Providers:**
    * Google's Gemini, Anthropic's Claude, Microsoft's Co-Pilot, XAI's Grok, DeepSeek (Chinese), Mistral Le Chat(French), and more.
- **Pricing Tiers & Tradeoffs:**
  - **Free Tiers:**
    * Usually include smaller models (like GPT-4 Mini) that might be less creative and more prone to inaccuracies.
  - **Paid Plans:**
    * Provide access to larger, more powerful models. For instance, ChatGPT Plus (around $20/month) offers GPT-4o with certain usage limits, while higher plans (such as Pro at $200/month) unlock additional features and offer unlimited access.
  - **Decision Factor:**
    * Pick the plan that best fits your professional requirements or casual use.

2.2 Thinking (Reasoning) Models

- **Definition:**
  - These models have been further refined using reinforcement learning to handle multi-step reasoning and tackle complex problem-solving.
- **Use Cases:**
  - They're especially handy for challenging tasks like advanced mathematics, debugging code, and other scenarios that demand deep logical thinking.
- **Tradeoffs:**
  - Because they "think" by generating extra internal tokens, they might take a bit longer—sometimes even minutes—to produce a response.
- **Example:**
  - When debugging a tricky programming issue (say, a gradient check failure), you might need to switch to a "thinking" model that carefully works through its reasoning steps before arriving at the solution.

---

3. Tool Integration and Advanced Functionalities

3.1 Internet Search and Deep Research

- **Internet Search Integration:**
  - **Function:**
    - It automatically spots when a query needs up-to-date or external information and kicks off a web search.
  - **How It Works:**
    - The model emits a special "search token" that tells the system to pull in and insert web content (with proper citations) into the conversation.
  - *Example:*
    - Asking about the release date of a new TV show or checking out the latest trends.
- **Deep Research:**
  - **Combination of Tools:**
    - It blends extended reasoning with multiple internet searches.
  - **Outcome:**
    - You get detailed reports—similar to a custom research paper—that include citations and references.
  - *Example:*
    - Investigating the properties and safety of AKG in a longevity supplement.

3.2 File Uploads and Document Analysis

- **Capabilities:**
  - You can upload PDFs, images, or text files so the model can reference and analyze real documents.
- **Practical Applications:**
  - Summarizing academic papers or books (like *The Wealth of Nations*).
  - Offering detailed interpretations of medical reports (say, a 20-page blood test PDF).

3.3 Python Interpreter and Code Generation

- **Integration with Code:**
  - The model is capable of writing and executing code (whether it's Python, JavaScript, etc.) when you need more than just a theoretical answer.
- **Process:**
  - It generates code, uses special tokens to run it, and then shows you the output as text.
- **Use Cases:**
  - Creating plots, analyzing trends, or troubleshooting code.
- **Best Practice:**
  - Always double-check the generated code for any hidden assumptions or potential errors.

3.4 Custom GPTs and Custom Instructions

- **Custom GPTs:**
  - **Definition:**
    - These are pre-configured prompts designed to handle tasks you do repeatedly (like language translation or vocabulary extraction).
  - **Benefit:**
    - They save you time by ensuring you get consistent and accurate output every time.
- **Custom Instructions:**
  - **Usage:**
    - This feature lets you set a global behavior, tone, and style for the model.

- *Example:*
    - You might tell the model to adopt a specific tone when translating Korean, such as maintaining a certain level of formality.

---

4. Multimodal Interactions

4.1 Audio and Voice

- **Speech-to-Text and Text-to-Speech:**
    - **Input:**
        - Converts your spoken queries into text—perfect for mobile use.
    - **Output:**
        - Reads the responses aloud so you can interact hands-free.
- **Advanced Voice Mode (True Audio):**
    - **Definition:**
        - This mode lets the model process and generate audio directly without converting to text first.
    - **Customization:**
        - You can tweak the model to speak in various styles—be it like Yoda, a pirate, or with a touch of romance.

4.2 Images and Video

- **Image Processing:**
    - **Method:**
        - Images get split into a grid of patches, and each patch is transformed into tokens.
    - **Applications:**
        - This is useful for generating creative images (using tools like DALL·E or ideogram) or analyzing visual data such as nutrition labels.
- **Video Analysis:**
    - **Emerging Capability:**
        - Some mobile apps now allow live video or sequential image analysis, enabling the model to "see" and interpret moving content.

---

5. Quality of Life and Customization Features

5.1 Persistent Memory

- **Memory Feature:**
  - The model can remember and recall personal details (like past chats or your preferences) stored in a dedicated "memory bank."
- **Benefits:**
  - This enhances personalization (for example, offering better movie recommendations).
  - The saved memory gets added to new conversations, keeping context consistent over multiple sessions.
- **Tip:**
  - Tell the model to "remember" key details to make future interactions smoother.

5.2 Custom Instructions

- **Customization:**
  - You can adjust how the model speaks—its tone and style. For instance, you might say, "don't be like an HR business partner—just speak normally and be educational."
  - This setting affects every conversation you have.

5.3 Custom GPTs for Specific Tasks

- **Definition:**
  - Custom GPTs are tailored configurations meant to handle tasks you do over and over again, like translating languages or pulling out vocabulary.
- **Implementation:**
  - They use few-shot prompting (providing several examples) to clearly show the model what kind of output you expect.
- **Example:**
  - A custom "Korean Vocabulary Extractor" that transforms sentences into flashcard-ready entries.

---

6. Practical Use Cases and Best Practices

6.1 Everyday and Professional Applications

- **Factual Query Handling:**
  - LLMs can be your go-to for everyday questions—whether you're wondering about the caffeine content in your drink or need travel advice.
- **Switching Contexts:**
  - When you change topics, it's best to start a new conversation to keep the context window clear and efficient.

6.2 Technical and Research Applications

- **Coding and Debugging:**
  - For those tough programming challenges, switch to a "thinking" model that uses reinforcement learning for deeper reasoning.
- **Data Analysis and Visualization:**
  - Use the integrated Python interpreter to write code that plots figures, extrapolates trends, and returns meaningful results.
- **Document Analysis:**
  - Upload academic papers or lengthy reports to get summaries and ask in-depth questions.

6.3 Multimodal and Interactive Research

- **Voice-Enabled Interaction:**
  - On mobile devices, take advantage of speech-to-text and text-to-speech features for quicker, more natural queries.
- **Image and Video Inputs:**
  - Native image tokenization and video analysis let you interact with rich, visual data seamlessly.
- **Custom Tools:**
  - Experiment with tools like "Deep Research" to generate detailed reports complete with citations.
- **Practical Tip:**
  - Always cross-check the outputs—especially when using deep research or code generation—with primary sources or a manual review.

7. Summary & Key Takeaways

- **Core Mechanisms:**
  - LLMs operate by splitting text into tokens within a limited context window and are built on two key training phases: pre-training (compressing knowledge) and post-training (adding personality and style).
- **Ecosystem Awareness:**
  - With various providers and pricing tiers, you can choose the model that fits your needs—whether you require a fast, compact model for casual queries or a larger, reasoning-enhanced model for complex challenges.
- **Tool Integration:**
  - Integrated tools like internet search, file uploads, and code interpreters are there to help enhance the precision and relevance of the model's responses.
- **Multimodal Interactions:**
  - Voice, image, and video functionalities offer more natural, engaging ways to interact with the model.
- **Customization:**
  - Leverage features like persistent memory, custom instructions, and custom GPTs to tailor the model to your specific tasks and personal style.
- **Best Practices:**
  - Keep the context clean by starting new chats when shifting topics.
  - Always review outputs—especially code and in-depth research—for accuracy.
  - Try out different tools to see which ones best suit your workflow.