Reading Assignment 6: Training LLMs

When the GPT-2 paper says *"the supervised objective is the same as the unsupervised objective, but only evaluated on a subset of the sequence,"* it's describing a profound unification between how language models learn and how they can perform tasks like translation, summarization, or question answering, without ever being explicitly told to do so. In supervised learning, we train on labeled pairs (input → output), like ("Translate to French: cat", "chat"), reducing the loss function. In unsupervised language modeling, the model simply predicts the next token in a sequence, (p(s_n | s_1, …, s_{n-1})). The insight is that these are *the same thing* when you write the supervised example as a single long sequence: "Translate to French: cat = chat"

During training, the model is predicting each next word (including "chat") as part of the sequence. The "unsupervised" objective (predict the next token) is, therefore, implicitly learning the "supervised" mapping (input → output) when the input and output appear together in natural text. The only difference is that in supervised training, we evaluate only the part corresponding to the output tokens. That's the subset the quote speaks of.

From a technical standpoint, this means that GPT-2 is optimizing a single probabilistic model $\big(p(\text{output}|\text{input})\big)$ *within* the general $objective(p(x) = \prod p(s_i|s_{<i}))$. There is no special rule or architecture for tasks,the model learns them from the distributional structure of language itself. T5 later adopt the same philosophy, treating every NLP task as "text-to-text": turning summarization, translation, and classification all into conditional next-token prediction problems.

We can see this idea at work today in models like GPT-5 and Claude, which perform question answering, coding, reasoning, and summarization without task-specific fine-tuning. When you type *"Translate this sentence to German:"*, the model is just continuing the sequence in the most probable way, guided by the statistical structure of billions of similar textual demonstrations it has absorbed. Here's an analogy. Imagine teaching a student entirely by letting them finish your sentences. If you say, "Translate to French: the sky is…", and they complete it as "le ciel est…", they've learned the task without any labeled "training data." The act of predicting what naturally comes next across enough varied examples is itself the process of learning.