Procesamiento de Lenguaje Natural Avanzado

Large Language Models (LLMs)

Evolution, Capabilities, and Limitations

Dra. Helena Gómez Adorno
helena.gomez@iimas.unam.mx

Dr. Fazlourrahman Balouchzahi
fbalouc@iimas.unam.mx

Correo del curso:
pln.cienciadedatos@gmail.com

A **Large Language Model (LLM)** is a neural network trained to predict the next token in a sequence, using massive amounts of text data.

LLMs are:

- Transformer-based architectures
- Trained with self-supervised learning
- Scaled to billions of parameters
- Trained on web-scale corpora

They predict the next token given previous tokens. That is the entire training objective.

- For example, given: "The capital of France is"
  - The model assigns high probability to "Paris" because that continuation was statistically reinforced during training.
- But now consider: "The capital of Wakanda is"
  - The model may still generate a fluent answer — even though Wakanda is fictional. The objective is to produce the most probable continuation, not to verify ontological truth.
- LLMs optimize likelihood, not epistemic validity.

What Is a Large Language Model?

When model size, data, and compute increase, performance improves predictably. But beyond certain scales, qualitative behavior appears.

- For example, small models struggle with few-shot learning:
  - `Text: "I feel hopeful about tomorrow."`
  - `Label: Hope`
  - `Text: "Nothing will ever improve."`
  - `Label: Hopelessness`
  - `Text: "I can't wait for next year."`
  - `Label: ??`

Large models infer the pattern and output "Hope" — without parameter updates.

This suggests that large models internalize patterns of task structure, not just language.

Is this reasoning? Or statistical extrapolation at scale?

LLMs are not symbolic reasoning engines.

They are high-dimensional statistical systems that approximate reasoning through learned distributional patterns.

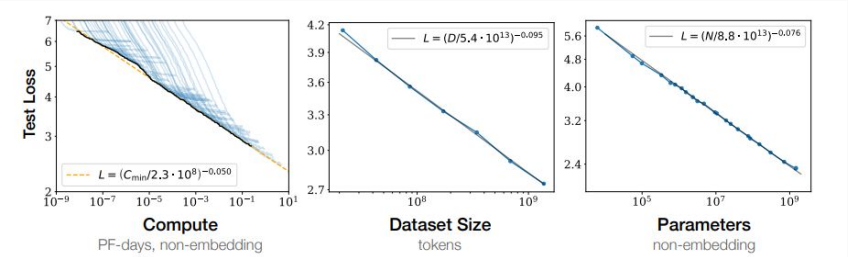

Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

Scaling and Behavioral Shifts

Inside the model, text becomes vectors in high-dimensional space. Attention reshapes these vectors based on context.

For example, The word "bank" in:

- `"She sat by the river bank."`

And in

- `"She deposited money in the bank."`

produces different contextual embeddings. The geometry of representation shifts based on surrounding tokens.

The model does not store dictionary meanings. It adjusts vector positions according to distributional constraints.

Meaning emerges from geometry.

Internal Representations Are Geometric

Consider this prompt:

```
- Translate English to Spanish:
  Dog → Perro
  Cat → Gato
  House →
```

The model outputs "`Casa`." No weights changed. The prompt conditioned the distribution.

Now increase complexity:

```
- Translate English to Spanish in a sarcastic tone:
- Dog → Perro (obviously)
- Cat → Gato (as if you didn't know)
- House →
```

The model adapts style.

This demonstrates that LLMs infer latent task patterns from prompt structure. It is meta-learning embedded during pretraining.

In-Context Learning: Apparent Learning Without Updates

Ask:

- `“Give me three academic citations about the 2025 Theory of Hope Optimization by Smith & Zhang.”`

The model may produce:

- `Smith, J., & Zhang, L. (2025). Hope Optimization Framework. Journal of Cognitive AI...`
- These references may not exist.
- The output is structurally correct — authors, year, journal format — but factually fabricated.
- The model learned the distribution of academic writing, not the existence of specific papers.
- Fluency can mask falsity.

Hallucinations: Fluent but False

Consider a simple reasoning task:

- "A train travels 60 km/h for 3 hours. How far does it travel?"

Correct answer: 180 km.

Now increase complexity:

- A train travels 60 km/h for 3 hours, then 80 km/h for 2 hours. What is the average speed?

- Some models produce incorrect arithmetic despite explaining steps confidently.
- This shows that LLM reasoning is often approximate sequence modeling, not symbolic computation.
- **Chain-of-thought** prompting improves performance — but does not guarantee correctness.

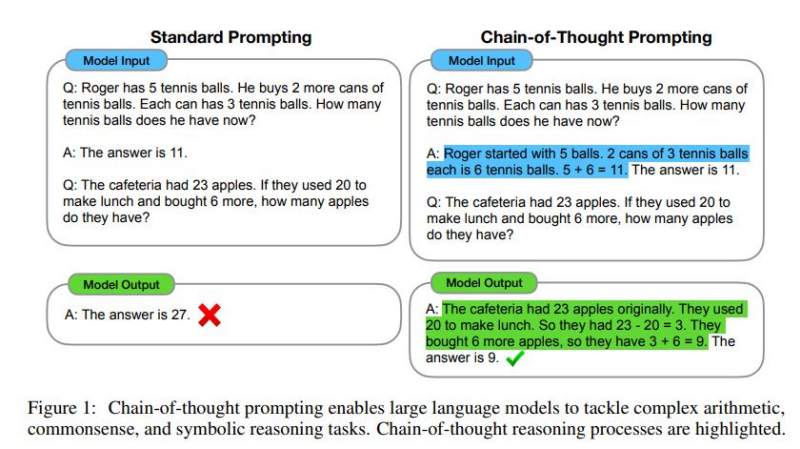

Figure 1: Chain-of-thought prompting enables large language models to tackle complex arithmetic, commonsense, and symbolic reasoning tasks. Chain-of-thought reasoning processes are highlighted.

Reasoning vs Pattern Completion

Suppose you give a 20-page document and ask:

- `“What was the main argument introduced in paragraph 2?”`

- If paragraph 2 falls outside the context window, the model cannot attend to it.
- The limitation is architectural: attention complexity grows quadratically with sequence length.
- Unlike humans, the model does not “remember” earlier pages unless they are explicitly in context or retrieved externally.
- Memory is bounded by token limits.

Context Window as a Cognitive Bottleneck

Ask:

- `"A nurse walked into the room. What did she do next?"`

The model might default to female pronouns.

Ask:

- `"A CEO walked into the room. What did he do next?"`

Gender associations may appear automatically.

These outputs reflect distributional statistics in training data. The model amplifies correlations it has observed.

Bias is not intentional. It is encoded in statistical structure.

Extreme *she* occupations

1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

Extreme *he* occupations

1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. figher pilot	12. boss

Bias as Statistical Reflection

LLMs perform well when tasks can be reduced to structured linguistic transformations.

For example:

- Summarizing a research abstract
- Rewriting text in formal tone
- Generating syntactically valid Python code
- Extracting entities from text

These tasks depend heavily on pattern regularities in language.

LLMs are extremely powerful linguistic simulators.

What LLMs Excel At?

They cannot:

- Guarantee factual correctness
- Verify sources internally
- Maintain persistent beliefs
- Access real-time reality without tools

If you ask:

```
"What is the temperature in Mexico City right now? "
```

Without external tools, the answer will be a plausible guess — not live data.

LLMs are generative statistical systems, not grounded agents.

What LLMs Cannot Reliably Do?

Evaluation Depends on What We Want the LLM To Do

There is no single “LLM score.” Evaluation depends on the intended use.

An LLM can function as:

- A language model (next-token predictor)
- A classifier
- A reasoning system
- A generator (summarizer, translator, coder)
- A conversational agent

Therefore, evaluation must be **task-aligned**.

1 As a Language Model

We evaluate predictive fit using **perplexity**:

Lower perplexity → better modeling of token distributions.

However:
Low perplexity ≠ factual correctness
Low perplexity ≠ reasoning ability

It measures statistical fit — not intelligence.

How do we evaluate a Large Language Model?

2 As a Task Solver

If the LLM performs classification, extraction, or structured prediction, we use classical ML metrics:

- Accuracy, Precision / Recall, Macro-F1, Exact match

Example:
If used for hope detection:

"I can't wait for tomorrow."

Gold label = Hope
We compute standard classification metrics.

When used as a task solver, evaluation mirrors supervised learning.

3 Generation Evaluation

For summarization or translation:

Metrics: ROUGE, BLEU, BERTScore

Challenge:
Multiple correct outputs exist.

Therefore:
Surface overlap ≠ semantic correctness.

Human evaluation is often required to judge coherence, relevance, and factuality.

How do we evaluate a Large Language Model?

- **Epistemic Limitations**
  - Hallucination
  - Approximate Reasoning
  - Lack of Grounding
- **Architectural Limitations**
  - Context Window Bound
  - Prompt Sensitivity
  - Computational Cost
- **Behavioral and Social Limitations**
  - Bias
  - Overconfidence
  - Inconsistency

Limitations

LLMs optimize next-token probability, not truth or logical validity.

- **Hallucination**

  Fluent but false outputs arise because the model maximizes likelihood, not factual correctness.

  Probability ≠ truth.

- **Approximate Reasoning**

  Reasoning performance is unstable and sensitive to phrasing.

  Small prompt changes can alter correctness.

- **Lack of Grounding**

  LLMs do not access real-time reality unless connected to tools.

  Their “knowledge” is compressed training data, not verified facts.

Epistemic Limitations

LLMs are constrained by their computational structure.

- **Context Window Bound**
  Attention is limited to a fixed token window.
  Information outside context is inaccessible.
- **Prompt Sensitivity**
  Behavior can change under minor paraphrasing.
  Reasoning is not fully invariant to surface form.
- **Computational Cost**
  Training and inference require large-scale compute.
  Performance must be balanced against efficiency and sustainability.

Architectural Limitations

LLMs inherit patterns from training data.

- **Bias**
  Statistical correlations in data lead to demographic asymmetries in outputs.
- **Overconfidence**
  Models may express high certainty even when incorrect.
  Calibration is imperfect.
- **Inconsistency**
  No persistent memory or stable belief system.
  Outputs can contradict across prompts.

Behavioral and Social Limitations

Large Language Models are large-scale probabilistic sequence models trained to predict the next token given context. Their power comes from scale — billions of parameters trained on massive corpora — which allows them to approximate complex linguistic and reasoning patterns.

What They Do Well

- Model language with high fluency
- Perform many NLP tasks via prompting
- Exhibit in-context learning
- Approximate multi-step reasoning
- Adapt style and structure dynamically

They are powerful **statistical simulators of language**.

Fundamental Limitations

- Optimize probability, not truth
- Hallucinate plausible but false content
- Reason approximately, not symbolically
- Limited by context window
- Sensitive to prompt phrasing
- Reflect biases in training data
- Computationally expensive

They generate **plausible continuations**, not verified knowledge.

LLMs at a Glance — Foundations, Capabilities, and Limits

Tarea 1 — Controlled Behavioral Analysis of an LLM

Title: **Analyzing the Behavioral Limits of a Large Language Model**

Instructions:

1. Choose one LLM (e.g., GPT, Claude, Gemini, LLaMA).
2. Design **5 controlled prompts**, each testing a different property:
    - One reasoning task
    - One hallucination test
    - One paraphrase robustness test
    - One bias-sensitive prompt
    - One factual knowledge question
3. For each case:
    - Provide the prompt
    - Provide the model output
    - Analyze the behavior
    - Classify the limitation (epistemic, architectural, social)

Deliverable:

A technical report.

Tarea 2 — Robustness Under Paraphrasing

Title: **Prompt Sensitivity and Stability in LLMs**

Instructions:

1. Select 5 factual or reasoning questions.
2. Create **3 paraphrased versions** of each question.
3. Query the same model with all variants.
   Measure:
   - Answer consistency
   - Accuracy differences
   - Confidence variation

Questions to answer:

- Does phrasing affect correctness?
- Is reasoning invariant to surface form?
- Which prompts produce instability?

Deliverable:

Short report + comparison table.