Predio Dr. Alberto Cassano - CCT CONICET Santa Fe
Colectora Ruta Nac. 168 km 0 – Paraje "El Pozo"
S3007ABA Santa Fe, Argentina
(+54 342) 4511370 int. 4001/4003
imal@santafe-conicet.gov.ar
www.imal.conicet.gov.ar

Instituto de Matemática
Aplicada del Litoral
"Dra. Eleonor Harboure"
CONICET - UNL



# SEMINARIO DEL IMAL 2025
## "Macías-Segovia"

## Decoding Semantic Ambiguity in Large Language Models: Aligning Human Behavioral Responses with GPT-2's Internal Representations

**Bruno Bianchi**

***Resumen.*** Large Language Models (LLMs) exhibit human-like text processing, yet their internal mechanisms for resolving semantic ambiguity remain opaque, similar to the "black box" of human cognition. This study investigates how LLMs disambiguate concrete nouns by comparing their semantic biases to human behavioral responses. A corpus of sentences containing ambiguous words (e.g., "arco") paired with biasing contexts (e.g., short paragraphs related to "football" and "architecture") was created. Human participants identified their perceived meanings of ambiguous words in these contexts, establishing a behavioral ground truth. To analyze GPT2's disambiguation processes, two technical steps were implemented: (1) the model was fine-tuned to obtain a word-based tokenization, and (2) each ambiguous word's meaning was defined using curated word lists. The model's semantic bias was measured via cosine distances between contextualized embeddings of ambiguous words. Results revealed that GPT2's middle layers correlated with human disambiguation patterns, particularly when using word-based tokenization, mirroring findings in human-model alignment research. This suggests shared computational principles between human cognition and LLM processing for resolving ambiguity. The study advances interpretability research by linking model-internal representations to human behavioral benchmarks, offering insights into both artificial and biological language systems.

***Bio.*** Dr. de la UBA en Cs Biológicas. Postdoc en el Laboratorio de Inteligencia Artificial Aplicada de Exactas-UBA. Profesor del Dpto de Computación (Exactas-UBA). Durante mi doctorado estudié los mecanismos cerebrales de las predicciones durante la lectura. Actualmente investigo la relación entre los mecanismos internos de los modelos de lenguaje y el cerebro humano.

## Viernes 25 de abril, 15:30 horas

El Seminario se realizará en la SUM del IMAL y se transmitirá por videoconferencia.

Los datos de conexión Zoom son los siguientes:

*ID de reunión:* 857 0845 9470

*Código de acceso:* aB0Wa=sˆ7=

**NOTA:** en algunos casos copiar y pegar el ID y el código no funciona para establecer la conexión. Probar tipear ambos.