Estado da publicação: Não informado pelo autor submissor

Proposta de framework formal: por uma semântica neurossimbólica híbrida

Daniel Perico Graciano

https://doi.org/10.1590/SciELOPreprints.13017

Submetido em: 2025-08-18
Postado em: 2025-09-11 (versão 1)
(AAAA-MM-DD)

Proposta de framework formal: por uma semântica neurossimbólica híbrida

A formal framework proposal: for a hybrid neurosymbolic semantics

Ensaio teórico

Daniel Perico Graciano[1]

Resumo: Com o avanço dos Modelos de Linguagem de Larga Escala (LLMs), como os populares chatbots, surge um novo conjunto de desafios e oportunidades para a semântica. Enquanto os LLMs demonstram uma capacidade impressionante de gerar e processar linguagem natural, a forma como eles representam e manipulam o significado abre um frutífero tópico de pesquisa ativa. A lacuna entre a semântica latente profunda dos LLMs e os formalismos simbólicos tradicionais da semântica formal é um problema em aberto. A questão de como extrair e formalizar o conhecimento semântico de LLMs, e como usar a semântica para controlar e explicar o comportamento desses modelos, é uma área emergente que sintetiza em si um problema em aberto.
Palavras-chave: Semântica. Modelos de Linguagem. Inteligência Artificial.

Abstract: With the advancement of Large Language Models (LLMs), such as popular chatbots, a new set of challenges and opportunities for formal semantics emerges. While LLMs demonstrate an impressive capacity to generate and process natural language, the way they represent and manipulate meaning opens up a fruitful topic of active research. The gap between the deep latent semantics of LLMs and traditional symbolic formalisms of formal semantics is an open problem. The question of how to extract and formalize semantic knowledge from LLMs, and how to use formal semantics to control and explain the behavior of these models, is an emerging area that itself synthesizes an open problem.
Keywords: Semantics. Language Models. Artificial Intelligence.

Resumo para não especialistas: Modelos de Linguagem de Larga Escala (LLMs), como os chatbots, são estruturas digitais muito grandes que aprendem a falar e escrever. Eles são muito bons em entender e criar textos, mas ainda estamos descobrindo como eles realmente "entendem" o significado das palavras. Esses LLMs têm um modo próprio de lidar com o significado, diferente da forma como os cientistas tradicionalmente organizam o significado das palavras. Este texto tenta responder duas perguntas importantes: Como podemos extrair e organizar o que os LLMs aprenderam sobre o significado das palavras? E como podemos

[1] Doutor em linguística, Programa de Pós-graduação em Linguística - Universidade Federal de São Carlos, São Carlos - São Paulo, e-mail: danip.graciano@gmail.com, ID ORCID: https://orcid.org/0000-0002-5269-0213

usar esse conhecimento para fazer com que os LLMs se comportem de forma mais previsível e para entender melhor por que eles dizem o que dizem?

Introdução

Considerando os avanços recentes e a direção da pesquisa em inteligência artificial e processamento de linguagem natural, um dos principais problemas em aberto na semântica formal pode ser identificado como a integração e harmonização da semântica formal tradicional com as representações de significado aprendidas por modelos de linguagem de larga escala (LLMs) (Lewis; Steedman, 2023). É inegável a eficácia dos LLMs no que concerne às tarefas de processamento de linguagem. No entanto, eles operam de modo que muitas vezes carece da interpretabilidade e do rigor formal que a semântica oferece. A ponte entre essas duas abordagens – a simbólica e a conexionista/distribucional – é fundamental para o futuro da compreensão da linguagem natural e para o desenvolvimento de sistemas de IA mais robustos e explicáveis. Esse problema abrange vários desafios, como a necessidade de formalizar o conhecimento semântico implícito nos LLMs, aprimorar a composicionalidade em modelos de aprendizado de máquina e desenvolver métodos para controlar semanticamente a geração de texto. A resolução dessa lacuna não apenas avançaria a teoria semântica, mas também teria implicações profundas para a inteligência artificial, permitindo a criação de sistemas que não apenas processam a linguagem, mas realmente a compreendem em um sentido formal e interpretável.

Zhang e Freitas (2024) abordam diretamente a desconexão crescente entre a representação de conhecimento dos modelos de linguagem de grande escala (LLMs) e a semântica tradicional. Eles destacam que, embora os LLMs demonstrem um desempenho impressionante em diversas tarefas de Processamento de Linguagem Natural (PLN), há uma dificuldade em capturar o conhecimento semântico de forma interpretável por humanos ou em explicar o comportamento de inferência do modelo. A pesquisa proposta no artigo visa encurtar a lacuna entre a semântica latente profunda e os formalismos simbólicos, focando na formalização do conhecimento, na extração de informações linguísticas de modelos distribucionais e em técnicas de intervenção para permitir raciocínio explicável e geração de texto controlável. Isso reforça a ideia de que a integração da semântica formal com os LLMs é um dos principais problemas em aberto, se não o principal, na área. O artigo também menciona a importância da composicionalidade na linguagem natural como um aspecto

crucial da pesquisa em representação de texto, e como a interpretabilidade e o controle são de crescente relevância para a comunidade de PLN como um todo. A necessidade de harmonizar a flexibilidade e a capacidade de entrega de tarefas dos grandes modelos distribucionais com a capacidade de rastrear seu conhecimento e comportamento em termos de propriedades formais bem definidas é apresentada como uma questão de pesquisa chave.

Neste artigo, propomos um framework formal para uma semântica neurossimbólica híbrida, no intuito de abrir uma discussão possível para o tratamento do problema supracitado, a começar por uma proposta de formalização da lacuna, seguida do esboço de uma ponte formal que permita a extração de representações simbólicas verificáveis a partir de representações distribuídas, e vice-versa, facilitando a inferência e o controle semântico.

1. Formalização da lacuna

A lacuna entre a Semântica Formal (SF) e os Modelos de Linguagem de Grande Escala (LLMs) pode ser formalmente definida como a discrepância entre a capacidade de representação e inferência de significado dos LLMs e os requisitos de interpretabilidade, composicionalidade e verificabilidade que são fundamentais, uma vez que os recentes avanços que postulam relações de escala de adjetivos e organização conceitual semelhante à de bases de conhecimento formais dependem de ajustes ou modelos maiores (c.f. Liu; Xiang; Ding, 2023), de modo que ambos mais se retroalimentam enquanto proposições teóricas que apontam respostas ao problema.

Seja L uma linguagem natural. O significado de uma expressão $\alpha \in L$ é mapeado para uma representação formal $[[\alpha]] \in M\,M$, onde é um modelo matemático (e.g., teoria de conjuntos, lógica de ordem superior) que permite inferências lógicas e a atribuição de condições de verdade. A composicionalidade é um princípio chave, onde $[[f(x)]] = F([[f]], [[x]])$ para funções f e argumentos x, e F é uma função de combinação semântica.

Em contraste, um LLM, denotado por M, opera sobre sequências de tokens $T = \langle t_1,t_2,\ldots,t_n\rangle$. O significado, para um LLM, é implicitamente codificado em seus parâmetros θ e manifestado através de distribuições de probabilidade sobre ′sequências de tokens, $P(T \mid T)$. A representação interna de significado em um LLM é um vetor de alta dimensão $v \in \mathbb{R}^d$, obtido através de funções de codificação E: $T \rightarrow v$. A inferência é realizada por operações sobre esses vetores no espaço latente. A lacuna surge porque não existe um mapeamento direto e transparente de v para M que permita a inspeção e compreensão humana das

condições de verdade ou das relações lógicas subjacentes. A representação de significado em LLMs é distribucional e opaca, enquanto na semântica é simbólica e transparente.

Além disso, os LLMs exibam um comportamento composicional em suas saídas, a função F que combina os significados das partes não é explicitamente definida ou acessível. É uma composicionalidade emergente e implícita, não uma composicionalidade estrutural e explícita. Em contrapartida, Chieng et al. (2025) mostram que LLMs exibem comportamentos composicionais robustos em tarefas específicas, e que a hierarquia funcional de camadas pode refletir níveis crescentes de abstração semântica, ainda que de forma não totalmente interpretável. A ausência de uma função composicional explícita não implica ausência de composicionalidade funcional. Vale ressaltar, no entanto, que a limitação especificidade dessas tarefas expõem por si só as fragilidades dos modelos. Além disso, quando se trata de métodos híbridos, nos quais LLMs são usados para gerar representações formais (como PDDL ou especificações matemáticas), eles apresentam apenas uma potencialidade de verificação posterior por sistemas simbólicos.

A ausência de um modelo M explícito nos LLMs impede a verificação formal de inferências e a garantia de ausência de inconsistências lógicas. A análise permite a prova de validade de argumentos e a identificação de contradições, o que é desafiador em LLMs sem um framework formal subjacente.

Formalmente, a lacuna é a ausência de uma função $G : R^d \rightarrow M$ tal que, para $\alpha \in$ `qualquer expressão` L $G(E(\alpha))$, seja semanticamente equivalente a $[[\alpha]]$, e que G M preserve as propriedades de composicionalidade e inferência lógica de

$G : \mathrm{Rd} \rightarrow \mathrm{M}$

Afinal, não há, até o momento, um método sistemático e transparente para mapear as representações vetoriais internas dos LLMs para estruturas formais que permitam inferência lógica e verificação, como exigido pela semântica formal. E ainda que LLMs apresentem comportamentos composicionais emergentes, a função de combinação semântica (F) não é explicitamente definida nem acessível, dificultando a preservação das propriedades formais. Assoma-se a isso o fato de a potencialidade de alguns LLMs gerar representações formais (por exemplo, PDDL), são barradas na correspondência entre o vetor latente e a estrutura formal indireta, a depender de prompting, i.e. elas não podem ser garantidas para todos os casos.

2. Proposta de framework formal: semântica neurossimbólica híbrida (SNH)

Propomos um framework de semântica neurossimbólica híbrida (SNH) que visa integrar as capacidades de representação distribuída dos LLMs com o rigor e a interpretabilidade da Semântica Formal. O objetivo é construir uma ponte formal G: $R^d \rightarrow M$ que permita a extração de representações simbólicas verificáveis a partir de representações distribuídas, e vice-versa, facilitando a inferência e o controle semântico. Para isso, é preciso considerar que frameworks híbridos frequentemente herdam problemas de escalabilidade dos sistemas simbólicos, tornando-se difíceis de aplicar em tarefas de larga escala ou domínios abertos. A flexibilidade dos LLMs pode ser reduzida quando subordinada a componentes simbólicos rígidos, limitando a generalização e a adaptação a novos contextos. A integração eficaz entre modelos neurais (como LLMs) e sistemas simbólicos exige abordagens inovadoras para garantir interpretabilidade, composicionalidade e verificabilidade. A literatura recente aponta caminhos concretos para solucionar os principais desafios, o que envolve superação dos desafios em semântica neurossimbólica híbrida passa por frameworks programáveis, técnicas de condicionamento semântico, validação simbólica e desenvolvimento de representações intermediárias, promovendo sistemas mais interpretáveis, verificáveis e generalizáveis.

Vejamos a seguir algumas propostas puramente teóricas, que, apesar de despretensiosas, abrem espaço para discussões mais robustas acerca do problema.

2.1. Componentes do framework

O framework SNH consiste em:

1. Um LLM pré-treinado, M, que mapeia sequências de tokens para vetores de alta dimensão no espaço latente. Formalmente, $E : L \rightarrow R^d$, onde $E(\alpha)$ é o embedding da expressão α.

2. Uma função ou conjunto de funções, $D : R^d \rightarrow S$, que decodifica os embeddings dos LLMs em representações simbólicas intermediárias S. Estas representações podem ser, por exemplo, grafos de conhecimento, estruturas lógicas de primeira ordem, ou representações semânticas abstratas que capturam predicados, argumentos e relações. O MDS pode ser implementado via técnicas de probing semântico, parsing neural-simbólico ou knowledge distillation.

3. Uma função ou conjunto de regras, $T : S \rightarrow M$, que traduz as representações

simbólicas intermediárias S para o modelo *M*. Este módulo garante a composicionalidade explícita e a interpretabilidade, mapeando elementos S de para entidades, propriedades e relações em *M*. Por exemplo, um predicado em S seria mapeado para uma função característica em *M*.

4. Um sistema de raciocínio formal, $V : M \rightarrow \{verdadeiro, falso, indefinido\}$, que opera sobre as representações em *M* para realizar inferências lógicas, verificar consistência e identificar contradições. Este módulo utiliza os axiomas e regras de inferência da lógica subjacente.

5. Uma função ou processo $Gc : M \rightarrow L$, que gera texto em linguagem natural a partir de representações formais em *M*, com a capacidade de controlar aspectos semânticos e pragmáticos da saída. Este módulo pode ser implementado via LLMs condicionados por representações simbólicas ou por técnicas de template-based generation.

Cabe notar que a estrutura apresentada para o SNH é detalhada em alinhamento com tendências atuais, mas a literatura aponta limitações recorrentes e desafios práticos que precisam ser endereçados para garantir viabilidade e impacto. A combinação de múltiplos módulos (LLM, decodificação simbólica, tradução, raciocínio formal, geração de linguagem), pode se tornar excessivamente complexa e, em decorrência disso, difícil de escalar para tarefas de grande porte ou domínios abertos. A integração de componentes neurais e simbólicos frequentemente resulta em sobrecarga computacional e desafios de desempenho. Além disso, a existência de funções D (decodificação) e T (tradução simbólica) gerais, robustas e interpretáveis ainda é um objetivo de pesquisa. Na prática, esses mapeamentos são frequentemente parciais, dependentes de domínio e difíceis de generalizar, especialmente para representações simbólicas complexas. No entanto, as arquiteturas modulares e técnicas de otimização (como execução simbólica em GPU) para mitigar problemas de escalabilidade e desempenho e os métodos de aprendizado adaptativo e program synthesis para que as funções D e T podem ser ajustadas automaticamente a diferentes domínios, reduzindo dependência de conhecimento manual. Desse modo, é possível adotar métricas objetivas para interpretabilidade e explicabilidade no intuito de facilitar a avaliação e comparação entre abordagens. Assoma-se a isso o desenvolvimento de métodos de V&V que cubram o sistema como um todo, incluindo testes automatizados de consistência lógica e robustez. Quando lidamos com frameworks neurossimbólicos recentes (como Dolphin ou Lobster) que já incorporam soluções para escalabilidade, integração eficiente e suporte a raciocínio simbólico em larga escala. Portanto, o principal obstáculo aqui é que os frameworks

SNH são promissores, mas exigem modularização, métricas claras, métodos automáticos de adaptação e validação integrada para superar limitações práticas e teóricas apontadas pela literatura recente.

2.2. Fluxo de operação

O fluxo de operação do framework SNH pode ser descrito da seguinte forma: dada uma expressão $\alpha \in$ L, o LLM gera M um embedding $E(\alpha) \in \mathbb{R}^d$. O MDS D decodifica $E(\alpha)$ em uma representação simbólica $s = D(E(\alpha)) \in$ S. O MMF T traduz s para uma representação formal $m = T(s) \in$ M. De modo que o MVI V opera sobre para realizar inferências, verificar consistência e derivar novas informações $'m' \in M$. O MGC gera uma expressão em linguagem natural $\alpha' = G_C(m') \in L$ a partir da representação formal m'.

É preciso atentar para o fato de que o pipeline sequencial pode propagar e amplificar erros, como imprecisões no embedding E(α) afetam D, que por sua vez impacta T e V. Isso pode comprometer a confiabilidade das inferências e da geração de linguagem natural, especialmente em domínios abertos ou com dados ruidosos e o fluxo é essencialmente unidirecional, daí a relevância de loops de feedback entre módulos (por exemplo, entre raciocínio formal e embeddings neurais) que permitam aprendizado contínuo, adaptação e correção de erros em tempo real. Afinal, ainda que o módulo V realize inferências formais, a explicabilidade e a validação do sistema como um todo ainda representam importantes desafios, especialmente quando as decisões dependem de múltiplos módulos com diferentes paradigmas de representação.

2.3. Formalização da integração

A integração pode ser vista como a construção de um homomorfismo aproximado entre o espaço latente dos LLMs e o modelo formal. Seja L o conjunto de sentenças da linguagem natural, o espaço de embeddings do LLM, e M o domínio semântico formal.

Definimos a função de significado formal $[[\cdot]] : L \rightarrow M$. Definimos a função de codificação do LLM $E : L \rightarrow R^d$. O espaço de embeddings dos LLMs não é um único espaço vetorial homogêneo, mas sim uma coleção de submanifolds estratificados, cada um com diferentes dimensões e propriedades locais. Isso dificulta a definição de um homomorfismo global simples e estável entre embeddings e representações formais (c.f. Li; Sarwzte, 2025).

Por isso, cabe considerar o mapeamento como um homomorfismo local ou estratificado, reconhecendo que a correspondência pode variar conforme o domínio semântico ou o tipo de sentença.

O objetivo é encontrar uma função de decodificação semântica $G : R^d \rightarrow M$ tal que, $\alpha \in L$ $G(E(\alpha))$ para qualquer , seja semanticamente equivalente a *[[α]]*. Isso implica G que deve preservar as relações semânticas e lógicas presentes em M. Mais precisamente, para um conjunto de operadores semânticos

$$O_M = \{O_{M,1}, O_{M,2}, \ldots\} \text{ M}$$

em (e.g., negação, conjunção, quantificação) e um conjunto correspondente de transformações $O_E = \{O_{E,1}, O_{E,2}, \ldots\}$ no espaço de embeddings (e.g., operações vetoriais), buscamos que para cada $O_{M,i} \in O_M$ e $O_{E,i} \in O_E$, a seguinte propriedade seja aproximadamente satisfeita:

$$G(O_{E,i}(E(\alpha_1), E(\alpha_2), \cdots)) \approx O_{M,i}(G(E(\alpha_1)), G(E(\alpha_2)), \ldots)$$

Esta é uma condição de composicionalidade aproximada no espaço de embeddings, onde as operações semânticas formais são refletidas por transformações no espaço latente. A função G é construída através da composição do MDS e do MMF: $G = T \circ D$. De modo que o treinamento do MDS e do MMF pode ser realizado através de técnicas de aprendizado supervisionado, onde pares (embedding, representação simbólica) são utilizados. A avaliação da qualidade de G envolveria métricas de fidelidade semântica e de preservação de inferências lógicas. Este framework permite que os LLMs atuem como "oráculos" de significado distribuído, enquanto a análise fornece a estrutura para interpretar, raciocinar e controlar esse significado de forma explícita e verificável. A lacuna é, portanto, preenchida pela criação de interfaces formais entre as representações distribuídas e simbólicas.

O problema é que a correspondência entre operadores semânticos (OM) e operações vetoriais (OE) no espaço de embeddings é, na prática, apenas aproximada e altamente sensível ao contexto. O espaço latente dos LLMs não foi projetado para refletir diretamente operações lógicas como negação, conjunção ou quantificação, tornando a preservação sistemática dessas relações um desafio em aberto. No entanto, arquiteturas que modelam

explicitamente relações semânticas e lógicas, como Graph Neural Networks (GNNs) e módulos de agregação contextual, conseguem capturar dependências e composições entre entidades e operadores. Além de que, dados sintéticos, transfer learning e adaptação de domínio ampliam a cobertura semântica e estrutural do treinamento. Considere um grafo semântico (G = (V, E)), onde:

(V) são entidades (ex: pessoas, objetos, conceitos)

(E) são relações (ex: "amigo de", "possui", "é parte de")

Entidades:

(v_1): "Maria"

(v_2): "João"

(v_3): "Cachorro"

Relações:

(e_1): ("Maria", "amigo de", "João")

(e_2): ("João", "possui", "Cachorro")

Seja (v) um nó do grafo, (h_v{(l)}) o embedding do nó (v) na camada (l), e ({N}(v)) o conjunto de vizinhos de (v):

[h_v{(l+1)} = \Σ\ (W^{(l)}{AGG}\({h_u{(l)} : {N}(v) } \ { h_v{(l)}}]

Onde:

({N}(v)): vizinhos de (v)

({AGG}): função de agregação (ex: soma, média, atenção)

(W^{(l)}): matriz de pesos treinável da camada (l)

(Σ): função de ativação (ex: ReLU)

Exemplos como (Ana, amigo de, "Carlos") potencializam a diversidade semântica durante o treinamento, ao passo que o treinamento do GNN em um domínio (ex: rede social) o adapta para outro (ex: rede de citações). Além disso, os pesos do GNN para capturar novas relações e entidades, mantendo a capacidade de generalização. Se o modelo aprende a inferir "X é amigo de quem possui Y" em dados sintéticos, pode transferir esse conhecimento para frases reais como:

"Pedro é amigo de quem possui um gato."

Como queremos demonstrar, os GNNs e módulos de agregação contextual permitem capturar e compor relações semânticas e lógicas de forma explícita, e técnicas como dados sintéticos e transfer learning ampliam a cobertura e a robustez do treinamento. Consideraremos ainda a seguinte sentença em linguagem natural:

(α): "*Todos os pássaros voam e Piu-Piu é um pássaro*"

O LLM recebe a expressão α e gera um embedding $E(\alpha) \in R^d$. O LLM, testado em vastos corpora de texto, aprende a mapear sequências de palavras para representações vetoriais densas que capturam relações semânticas e sintáticas implícitas. Embora a natureza exata desses embeddings seja opaca, a hipótese é que eles codificam informações suficientes para a extração de significado. A função E é uma transformação complexa (e.g., rede neural profunda) que projeta a sentença em um espaço vetorial.

Seja α = "*Todos os pássaros voam e Piu-Piu é um pássaro*"

O LLM *M* computa (α): "*Todos os pássaros voam e Piu-Piu é um pássaro*", onde E(α) = vα', onde vα $\in R^d$ é o vetor de embedding correspondente à sentença.

O MDS *D* decodifica o embedding em uma representação simbólica intermediária *s* = $D(v\alpha) \in S$. Ele cria uma ponte entre o espaço contínuo e opaco dos embeddings e um espaço discreto e estruturado de representações simbólicas. Além de ser potencialmente implementado, por exemplo, como uma rede neural treinada que permite prever uma estrutura lógica a partir do embedding. Para este exemplo, o MDS é capaz de identificar os quantificadores, predicados e constantes presentes na sentença e suas relações.

O MDS transforma *D vα* em *s*α $\in S$. Seja um conjunto de fórmulas de lógica de primeira ordem. Idealmente, *sα* seria: $s\alpha = \{\forall x(Pássaro(x) \rightarrow V\ oa(x)), Pássaro(Piu - Piu)\}$. O MDS deve ser capaz de extrair a estrutura lógica subjacente. A primeira parte da sentença, "Todos os pássaros voam", é uma generalização universal. A segunda parte, "Piu-Piu é um pássaro", é uma afirmação de instanciamento. O MDS, através de treinamento em pares (sentença, fórmula lógica), aprende a reconhecer esses padrões.

O MMF *T* traduz a representação simbólica *sα* para uma representação formal $m = T(s\alpha) \in \mathbf{M}$. O MMF garante que a representação simbólica seja mapeada para um modelo

formal bem definido, onde as condições de verdade e as regras de inferência são explicitamente estabelecidas. Para este exemplo, pode ser um modelo de lógica de primeira ordem, onde predicados são conjuntos e constantes são elementos. *T* transforma *s*α *m*α ∈ *M* *M* = ⟨*D, I*⟩ um modelo de lógica de primeira ordem, onde é o domínio e é a função de interpretação. Para *s*α = {∀x(Pássaro(x) → Voa(x)), Pássaro(Piu – Piu)}:

I (Pássaro) ⊆ *D* (o conjunto de todos os pássaros no domínio)

I (Voa) ⊆ *D* (o conjunto das entidades que voam no domínio)

I (Piu – Piu) ∈ D (o indivíduo Piu–Piu no domínio)

A primeira fórmula, ∀*x(Pássaro(x) → Voa(x))*, é verdadeira em *M* se e somente se para todo *d* ∈ *D*, se *d* ∈ *I* (Pássaro), então *d* ∈ *I* (Voa). Isso significa que o conjunto de pássaros é um subconjunto do conjunto de entidades que voam *I (Pássaro)* ⊆ *I (Voa)*. A segunda fórmula (Pássaro(Piu – Piu)) é verdadeira em *M* se e somente se *I (Piu – Piu)* ∈ *I (Pássaro)*. O MMF estabelece a semântica formal das sentenças. Ele define como os símbolos da lógica de primeira ordem se relacionam com o mundo (o domínio D) e suas propriedades (a função de interpretação I). Isso é crucial para a verificação de inferências.

O MVI *V* opera sobre *m*α para realizar inferências e derivar novas informações *m'* ∈ *M*, por meio de regras de inferência da lógica de primeira ordem para deduzir novas verdades a partir das premissas. Neste caso, a regra de *modus ponens* generalizada é aplicável.

Dado: *m*α = {∀ *x(Pássaro(x) → Voa(x)), Pássaro(Piu – Piu)}*

1. ∀*x(Pássaro(x) → Voa(x))* (premissa 1)
2. *(Pássaro(Piu – Piu))* (Premissa 2)

Por instanciação universal (de 1), podemos inferir: 3. *Pássaro(Piu – Piu) → Voa(Piu – Piu)*.

Por *modus ponens* (de 2 e 3), podemos inferir: 4. *Voa(Piu – Piu)*.

Assim, *m'* = *{Voa(Piu – Piu)}* é a nova informação derivada, onde Voa(Piu – Piu) é verdadeira em *M* se e somente se *I (Piu – Piu)* ∈ *I (Voa)*. A inferência é logicamente válida se todos os pássaros voam e Piu-Piu é um pássaro, então Piu-Piu deve voar. O MVI formaliza esse processo dedutivo, garantindo a correção da inferência.

Deve-se levar em conta que as LLMs são capazes de mapear sentenças como “Todos os pássaros voam e Piu-Piu é um pássaro” para embeddings vetoriais densos, no entanto, a natureza opaca desses embeddings dificulta a extração direta de estruturas lógicas precisas. O processo de decodificação semântica, que busca transformar esses vetores em representações simbólicas intermediárias, depende fortemente do treinamento supervisionado com pares de sentenças e fórmulas lógicas, mas ainda enfrenta limitações na identificação de quantificadores, predicados e relações complexas, especialmente quando a sentença exige múltiplos passos de raciocínio ou envolve cadeias longas de dependências lógicas. Além disso, a coerência lógica das representações extraídas nem sempre é garantida, uma vez que os LLMs tendem a apresentar desempenho inferior ao humano em tarefas que exigem dedução formal rigorosa, frequentemente confundindo generalizações com quantificações universais e sendo sensíveis a pequenas variações na formulação textual (Lin et al., 2025). Outro ponto crítico é a escalabilidade, dado que à medida que a complexidade lógica das tarefas se intensifica, observa-se uma queda acentuada na precisão dos modelos, fenômeno conhecido como “maldição da complexidade”, que persiste mesmo com modelos maiores e mais recursos computacionais (Lin et al., 2025).

Por outro lado, a integração de LLMs com solucionadores simbólicos formais, como é o caso dos motores de inferência lógica, pode aumentar a precisão e a confiabilidade das deduções, permitindo que o LLM atue como tradutor semântico enquanto o sistema simbólico garante a validade formal das inferências. O aprimoramento do treinamento supervisionado, com conjuntos extensos de exemplos pareados entre linguagem natural e lógica formal, contribui para que o MDS reconheça padrões lógicos com maior fidelidade. Estratégias de raciocínio em cadeia, como o chain-of-thought, estimulam o modelo a explicitar cada etapa do raciocínio, melhorando a coerência e a capacidade de resolver problemas complexos. Além disso, abordagens como Best-of-N sampling, backtracking e prompts de auto-verificação têm se mostrado eficazes para aumentar a robustez e a precisão das inferências lógicas, especialmente em cenários de alta complexidade (Lin et al., 2025). Por fim, avanços em técnicas de compressão, paralelismo e aceleração de inferência, como decodificação especulativa e métodos de saída antecipada, contribuem para tornar o processo mais eficiente sem comprometer a qualidade das representações lógicas geradas.

Voltando a nossa sentença, MGC *Gc* gera uma expressão em linguagem natural $\alpha' = Gc\ (m') \in L$, a partir da representação formal *m’*, já que traduz a representação formal de volta para a linguagem natural. Ele deve ser capaz de gerar sentenças gramaticalmente

corretas e semanticamente equivalentes à representação formal. Isso pode ser feito por um LLM condicionado pela estrutura lógica ou por um sistema de geração baseado em regras. Portanto, MGC *GC m´ = {V oa(Piu − Piu)} α ∈´ L. Ao passo que as possíveis saídas para α'* são*: * "Piu-Piu voa" * "Piu-Piu é capaz de voar"*. A saída em linguagem natural reflete a inferência lógica realizada. O MGC demonstra a capacidade do framework de não apenas compreender e raciocinar sobre o significado, mas também de expressar os resultados de forma inteligível para humanos. Este exemplo demonstra como o framework SNH preenche a lacuna entre a semântica formal e os LLMs. A informação da linguagem natural é codificada, decodificada em uma representação simbólica, formalizada, raciocinada sobre ela e, finalmente, gerada de volta para a linguagem natural. Cada etapa é justificada por argumentos lógicos e formalizações, garantindo a coerência e a verificabilidade do processo. O SNH permite que a opacidade dos LLMs seja mitigada pela interpretabilidade e rigor da semântica formal, abrindo caminho para sistemas de IA mais robustos e confiáveis.

A geração de linguagem natural a partir de representações formais, como realizada pelo módulo gerador de comunicação (MGC) no framework SNH, é um processo fundamental para tornar inferências lógicas acessíveis e compreensíveis para humanos. O objetivo do MGC é traduzir uma expressão formal, por exemplo, em uma sentença em linguagem natural gramaticalmente correta e semanticamente equivalente, como "Piu-Piu voa" ou "Piu-Piu é capaz de voar". Essa tarefa pode ser realizada por grandes modelos de linguagem condicionados por estruturas lógicas ou por sistemas baseados em regras. De acordo com Li, Tian e Ji (2024), abordagens que consideram a estrutura hierárquica das expressões lógicas, como modelos estruturais (por exemplo, SLEtoNL), superam significativamente os métodos sequenciais tradicionais e modelos de linguagem pré-treinados como T5, BART e GPT-3, em especial em cenários fora do domínio de treinamento. No entanto, desafios persistem, como a tendência de modelos neurais gerarem sentenças "alucinadas" (não fiéis à entrada formal) e a limitação de expressividade quando o modelo não captura corretamente todas as dependências semânticas.

Para mitigar esses problemas, mecanismos de atenção, planejamento de conteúdo e treinamento supervisionado com pares de dados (formal-natural) têm se mostrado eficazes para aumentar a fidelidade e a naturalidade das sentenças geradas (c.f. Wu et al., 2023), de modo que o framework SNH demonstra ser possível preencher a lacuna entre semântica formal e linguagem natural, promovendo interpretabilidade, verificabilidade e robustez em

sistemas de IA.

3.Verificação da coerência e consistência

A proposta de semântica neurossimbólica híbrida (SNH) busca a coerência e consistência em múltiplos níveis: lógico, linguístico, matemático e epistemológico. A verificação desses aspectos é crucial para a validade do framework. A consistência lógica do framework SNH depende primariamente da fidelidade do Módulo de Mapeamento Formal (MMF) e do Módulo de Verificação e Inferência (MVI). O MMF, $T : S \rightarrow M$, deve garantir que as representações simbólicas intermediárias S sejam traduzidas para M de forma a preservar as relações lógicas. Isso implica que se uma inferência é válida em M, a representação correspondente em S (e, idealmente, no espaço de embeddings) deve suportar essa inferência. O MVI que opera sobre M, é intrinsecamente projetado para manter a consistência lógica. Qualquer inconsistência detectada pelo MVI (e.g., uma contradição) seria um sinal de falha no mapeamento ou na representação simbólica gerada pelo MDS. A meta é que, se $m \in$ `M é uma representação formal consistente, então` $G(m)^{-1}$ (a representação no espaço de embeddings que a gerou) não deve levar a inferências inconsistentes quando processada pelo LLM e, em seguida, mapeada de volta para M.

O desafio reside na aproximação da composicionalidade no espaço de embeddings: $G(O_{E,i}(E(\alpha_1),\cdots)) \approx O_{M,i}(G(E(\alpha_1)),\ldots)$. A "aproximação" implica que pode haver desvios. A consistência lógica é mantida no nível de M, mas a propagação de inconsistências do LLM para M deve ser mitigada. Isso pode ser feito através de restrições no MDS e mecanismos de correção. O MDS pode ser treinado para priorizar a geração de representações simbólicas que são logicamente consistentes dentro de um subconjunto de S que é mapeável para M sem contradições. Enquanto o MVI pode não apenas detectar inconsistências, mas também sugerir correções para as representações em S ou R^d que as causaram, realimentando o sistema. A consistência linguística refere-se à capacidade do framework de capturar e preservar as nuances do significado em linguagem natural. A lacuna original destaca que os LLMs são opacos, enquanto a SF é transparente. O SNH busca tornar o significado dos LLMs mais transparente através do MDS e MMF. Lidamos com ambiguidade por meio de múltiplas representações formais para uma única expressão ambígua. O SNH pode estender isso

permitindo que o MDS gere um conjunto de representações simbólicas para uma expressão ambígua, cada uma correspondendo a uma interpretação possível. A vagueza, por sua vez, pode ser modelada em M usando lógicas fuzzy ou superavaliação, e o MDS precisaria ser capaz de extrair as informações necessárias dos embeddings para suportar essas representações. O MGC (Módulo de Geração Controlada) pode ser aprimorado para levar em conta o contexto pragmático ao gerar texto a partir de M. Para Lewis e Steedman, "o principal desafio reside em harmonizar a semântica formal, com sua transparência e rigor, com as representações opacas, mas contextualmente ricas, dos Modelos de Linguagem de Larga Escala (LLMs), a fim de lidar com a ambiguidade e a vagueza inerentes à linguagem natural" (Lewis; Steedman, 2023, p. 165). Esse problema envolve a incorporação de representações de contexto no modelo formal ou o uso de LLMs ajustados para gerar saídas contextualmente apropriadas com base nas representações formais.

Por fim, a dependência crítica da fidelidade do Módulo de Mapeamento Formal (MMF) e do Módulo de Verificação e Inferência (MVI) para assegurar a consistência lógica é um ponto vulnerável, que renderia uma pesquisa de grande relevância, posto que a tradução entre representações simbólicas intermediárias (S) e formais (M) é extremamente desafiadora, principalmente no que concerne a natureza probabilística e aproximada dos embeddings. Quanto à aproximação na preservação da composicionalidade no espaço de embeddings, encontra-se aí uma evidência de que os desvios podem ocorrer e provavelmente ocorrerão, o que coloca em risco a consistência lógica final do sistema quando inferências passam pelo LLM e retornam a M. Para mitigar esse problema, seria necessário, além dos mecanismos de correção derivados do MVI, estratégias robustas de verificação e aprendizagem contínua que detectem e aprendam com essas inconsistências de maneira dinâmica, além do ajuste fino dos módulos para minimizar desvios desde o início. Outro ponto ainda não solucionado é a a abordagem da consistência linguística, especificamente na tentativa de lidar com ambiguidade e vagueza por meio da geração múltipla de representações formais e do uso de lógicas fuzzy ou superavaliação. O desafio envolve tanto representar adequadamente a ambiguidade, como também integrar essas múltiplas interpretações em um processo inferencial coerente que não comprometa a usabilidade prática do sistema. Na ausência de um método claro para resolver conflitos ou escolher interpretações mais adequadas em contextos variados, o sistema pode produzir resultados indecisos ou incoerentes.

Além disso, reconhecendo que a opacidade dos LLMs dificulta a transparência do significado, a solução apresentada baseia-se na mediação pelo Módulo de Mapeamento Formal e pelo Módulo de Geração Controlada (MGC), que tornariam os significados mais

transparentes, pode ser insuficiente se os próprios LLMs continuarem gerando representações com ruídos semânticos (e/ou contextuais) que não sejam plenamente capturados pelo sistema simbólico. A incorporação efetiva do contexto pragmático, como sugerido, é um problema aberto e difícil, especialmente na medida em que o contexto envolve uma gama de contingências que tendem ao infinito, o que impossibilita a previsibilidade demandada. Para superar essa limitação, seria útil desenvolver métodos híbridos que combinem análise de contexto explícita, aprendizagem de representações contextuais com supervisão simbólica, e feedback iterativo entre os módulos simbólicos e os LLMs para refinamento adaptativo das representações.

Por fim, a argumentação poderia se beneficiar de uma maior clareza quanto à operacionalização prática do sistema, incluindo exemplos concretos de como os módulos interagem durante processos reais de inferência, geração e correção, assim como métricas claras para avaliar a coerência e a consistência em múltiplos níveis. Isso facilitaria a compreensão, a reprodutibilidade e o aprimoramento sistemático do framework SNH.

4. Considerações finais

A consistência lógica do framework reside na solidez das transformações entre os espaços. A função $G = T \circ D$ deve ser bem definida e as operações no espaço de embeddings devem ter propriedades matemáticas que permitam a aproximação da composicionalidade. Isso implica em algumas métricas de similaridade nas quais a definição de "semanticamente equivalente" para $G(E(\alpha)) \approx [[\alpha]]$ requer métricas de similaridade no espaço M que sejam compatíveis com as métricas de similaridade no espaço R^d.

Os LLMs adquirem conhecimento de forma estatística e distribucional, enquanto a semântica se baseia em princípios racionais e dedutivos. O framework permite a extração de conhecimento implícito dos LLMs (via MDS) e sua formalização em conhecimento explícito (via MMF e MVI). Isso torna o conhecimento dos LLMs mais auditável e justificável. Além disso, a combinação do raciocínio dedutivo (no MVI) com as capacidades inferenciais dos LLMs (no Módulo de Codificação Distribuída e MGC) permite um sistema lógico mais robusto, que pode lidar tanto com a lógica formal quanto com a flexibilidade da linguagem natural. De modo que a verificação da coerência e consistência do SNH é um processo contínuo de refinamento e validação em cada um desses níveis. A “aproximação” na composicionalidade é o ponto mais crítico, exigindo pesquisa contínua para minimizar os

desvios e garantir que o sistema híbrido seja robusto e confiável. A meta não é que os LLMs se tornem sistemas de lógica formal, mas que suas representações possam ser interpretadas e controladas por tais sistemas, aproveitando o melhor de ambos os mundos.

Referências

CARVALHO, D. S.; ZHANG, Y.; FREITAS, A. Formal Semantic Controls over Language Models. In **Proceedings of the 2024 Joint International Conference on Computational Linguistics**, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries, 2024, pages 50–55, Torino, Italia. ELRA and ICCL.

CHIENG, E; DIOMO, D; KARVADEC, C; MACOCCO, I; YU, J; LAIO, A; BARONI, M. Emergence of a high-dimensional abstraction phase in language transformers. **arXiv**: 2502.01100v1 [**ICLR]**, 24 mai 2025.

LEWIS, M.; STEEDMAN, M. Bridging Formal Semantics and Neural Language Models: Challenges and Prospects. **Annual Review of Linguistics**, v. 9, p. 157–181, 2023.

LI, X; SARWATE, A. D. Unraveling the Localized Latents: Learning Stratified Manifold Structures in LLM Embedding Space with Sparse Mixture-of-Experts. **Cs. LG**, 2025.

LI, X; Tian, Y; JI, S. Semantic- and relation-based graph neural network for knowledge graph completion. **Appl Intell 54**, 6085–6107 (2024). https://doi.org/10.1007/s10489-024-05482-2

LIN, B.Y; BRAS, R.L; RICHARDSON, K; SABHARWAL, A; POOVENDRAN, R; CLARCK, P; CHOI, Y. ZebraLogic: On the Scaling Limits of LLMs for Logical Reasoning. **arXiv**: 2502.01100v1 [cs.AI] 03 Feb 2025.

LIU, W; XIANG, M; DING, N. Adjective Scale Probe: Can Language Models Encode Formal SemanticsInformation? **The Thirty-Seventh AAAI Conference on Artificial Intelligence** (AAAI- 2023).

MÜLLER, A.; BORGES NETO, J.; OLIVEIRA, R. P. A semântica formal das línguas naturais: histórias e desafios. **Revista de Estudos da Linguagem,** 2012, 20(1), 119-148. Disponível em: https://periodicos hml.cecom.ufmg.br/index.php/relin/article/view/28635/22555

PARTEE, B. H. Formal Semantics: Origins, Issues, Early Impact. **Baltic International Yearbook of Cognition, Logic and Communication**, 2011, 6. Disponível em: https://newprairiepress.org/cgi/viewcontent.cgi?article=1056&context=biyclc

Stanford Encyclopedia of Philosophy. Situations in Natural Language Semantics.

Disponível em: https://plato.stanford.edu/entries/situations semantics/

O autor não tem conflitos de interesse a declarar.

Este preprint foi submetido sob as seguintes condições:

- Os autores declaram que os necessários Termos de Consentimento Livre e Esclarecido de participantes ou pacientes na pesquisa foram obtidos e estão descritos no manuscrito, quando aplicável.
- Os autores declaram que a elaboração do manuscrito seguiu as normas éticas de comunicação científica.
- Os autores declaram que estão cientes que são os únicos responsáveis pelo conteúdo do preprint e que o depósito no SciELO Preprints não significa nenhum compromisso de parte do SciELO, exceto sua preservação e disseminação.
- Os autores declaram que os dados, aplicativos e outros conteúdos subjacentes ao manuscrito estão referenciados.
- O manuscrito depositado está no formato PDF.
- Os autores declaram que a pesquisa que deu origem ao manuscrito seguiu as boas práticas éticas e que as necessárias aprovações de comitês de ética de pesquisa, quando aplicável, estão descritas no manuscrito.
- Os autores declaram que uma vez que um manuscrito é postado no servidor SciELO Preprints, o mesmo só poderá ser retirado mediante pedido à Secretaria Editorial do SciELO Preprints, que afixará um aviso de retratação no seu lugar.
- Os autores concordam que o manuscrito aprovado será disponibilizado sob licença Creative Commons CC-BY.
- O autor submissor declara que as contribuições de todos os autores e declaração de conflito de interesses estão incluídas de maneira explícita e em seções específicas do manuscrito.
- Os autores declaram que o manuscrito não foi depositado e/ou disponibilizado previamente em outro servidor de preprints ou publicado em um periódico.
- Caso o manuscrito esteja em processo de avaliação ou sendo preparado para publicação mas ainda não publicado por um periódico, os autores declaram que receberam autorização do periódico para realizar este depósito.
- O autor submissor declara que todos os autores do manuscrito concordam com a submissão ao SciELO Preprints.