big data and
cognitive computing

MDPI

Article

Semantic Reasoning Using Standard Attention-Based Models: An Application to Chronic Disease Literature

Yalbi Itzel Balderas-Martínez [1], José Armando Sánchez-Rojas [2,†], Arturo Téllez-Velázquez [2], Flavio Juárez Martínez [2], Raúl Cruz-Barbosa [2], Enrique Guzmán-Ramírez [2], Iván García-Pacheco [2] and Ignacio Arroyo-Fernández [2,*,†]

[1] Instituto Nacional de Enfermedades Respiratorias (INER) Ismael Cosío Villegas, Ciudad de México 14080, Mexico; yalbibalderas@gmail.com

[2] Graduate Studies Division, Universidad Tecnológica de la Mixteca, Huajuapan de León 69000, Mexico; asanchez7714@gmail.com (J.A.S.-R.); atellezv@gs.utm.mx (A.T.-V.); flaviojuarezmtz@gmail.com (F.J.M.); rcruz@gs.utm.mx (R.C.-B.); eguzman@gs.utm.mx (E.G.-R.); ivan@gs.utm.mx (I.G.-P.)

[*] Correspondence: iaf@gs.utm.mx; Tel.: +52-953-53-20399

[†] These authors contributed equally to this work.

Abstract: Large-language-model (LLM) APIs demonstrate impressive reasoning capabilities, but their size, cost, and closed weights limit the deployment of knowledge-aware AI within biomedical research groups. At the other extreme, standard attention-based neural language models (SANLMs)—including encoder–decoder architectures such as Transformers, Gated Recurrent Units (GRUs), and Long Short-Term Memory (LSTM) networks—are computationally inexpensive. However, their capacity for semantic reasoning in noisy, open-vocabulary knowledge bases (KBs) remains unquantified. Therefore, we investigate whether compact SANLMs can (i) reason over hybrid OpenIE-derived KBs that integrate commonsense, general-purpose, and non-communicable-disease (NCD) literature; (ii) operate effectively on commodity GPUs; and (iii) exhibit semantic coherence as assessed through manual linguistic inspection. To this end, we constructed four training KBs by integrating ConceptNet (600k triples), a 39k-triple general-purpose OpenIE set, and an 18.6k-triple OpenNCDKB extracted from 1200 PubMed abstracts. Encoder–decoder GRU, LSTM, and Transformer models (1–2 blocks) were trained to predict the object phrase given the subject + predicate. Beyond token-level cross-entropy, we introduced the Meaning-based Selectional-Preference Test (MSPT): for each withheld triple, we masked the object, generated a candidate, and measured its surplus cosine similarity over a random baseline using word embeddings, with significance assessed via a one-sided t-test. Hyperparameter sensitivity (311 GRU/168 LSTM runs) was analyzed, and qualitative frame–role diagnostics completed the evaluation. Our results showed that all SANLMs learned effectively from the point of view of the cross entropy loss. In addition, our MSPT provided meaningful semantic insights: for the GRUs (256-dim, 2048-unit, 1-layer): mean similarity (μ_{sts}) of 0.641 to the ground truth vs. 0.542 to the random baseline (gap 12.1%; $p < 10^{-180}$). For the 1-block Transformer: $\mu_{sts} = 0.551$ vs. 0.511 (gap 4%; $p < 10^{-25}$). While Transformers minimized loss and accuracy variance, GRUs captured finer selectional preferences. Both architectures trained within <24 GB GPU VRAM and produced linguistically acceptable, albeit over-generalized, biomedical assertions. Due to their observed performance, LSTM results were designated as baseline models for comparison. Therefore, properly tuned SANLMs can achieve statistically robust semantic reasoning over noisy, domain-specific KBs without reliance on massive LLMs. Their interpretability, minimal hardware footprint, and open weights promote equitable AI research, opening new avenues for automated NCD knowledge synthesis, surveillance, and decision support.

check for updates

Academic Editor: Domenico Ursino

Received: 24 March 2025
Revised: 3 June 2025
Accepted: 3 June 2025
Published: 19 June 2025

Citation: Balderas-Martínez, Y.I.; Sánchez-Rojas, J.A.; Téllez-Velázquez, A.; Juárez Martínez, F.; Cruz-Barbosa, R.; Guzmán-Ramírez, E.; García-Pacheco, I.; Arroyo-Fernández, I. Semantic Reasoning Using Standard Attention-Based Models: An Application to Chronic Disease Literature. *Big Data Cogn. Comput.* **2025**, *9*, 162. https://doi.org/10.3390/bdcc9060162

Keywords: semantic reasoning; common-sense knowledge; natural language processing; chronic diseases

1. Introduction

Biomedical knowledge is expanding at a rate that outstrips any human reader. Transforming this deluge of literature into actionable insight therefore requires automated systems that can build, represent, and—critically—reason over large-scale textual data [1]. Two obstacles persist: (i) constructing high-coverage yet affordable knowledge bases (KBs) and (ii) deploying language models that perform robust semantic reasoning under the tight computational budgets typical of clinical and public-health settings.

Conventional relation extraction pipelines and curated knowledge graphs achieve high precision but suffer from low recall and expensive schema design [2,3]. Open Information Extraction (OpenIE) offers a low-cost, high-throughput alternative, yet the resulting noisy triples challenge downstream reasoning tasks [4]. Large language models (LLMs) excel at complex prompting and reasoning strategies—e.g., Chain-of-Thought, Tree-of-Thought, Least-to-Most—yet their compute demands often exceed what most biomedical groups can support [5]. In contrast, Standard Attention-Based Neural Language Models (SANLMs), such as Attention-Based Gated Recurrent Unit (GRU), Long-Short Term Memory (LSTM) networks, and Transformers, can be trained comfortably on commodity GPUs, but their semantic-reasoning limits remain under-explored.

Consequently, the community lacks a lightweight, end-to-end pipeline that (a) builds domain-specific KBs at scale, (b) equips compact SANLMs with robust reasoning skills, and (c) evaluates semantic coherence without costly expert annotation.

We tackle this gap through three questions: How does combining a common sense KB with general-purpose and specialized OpenIE KBs affect the reasoning performance of SANLMs in both generic and non-communicable-disease (NCD) domains? Can a Semantic Textual Similarity (STS)-based metric capture semantic coherence more faithfully than token-level accuracy? How do attention-based GRUs, LSTMs, and Transformers differ in reasoning accuracy and semantic error patterns?

Contributions

- Framework integration: a pipeline that fuses commonsense knowledge with OpenIE-derived KBs centered on NCD literature, aligning large-scale text mining with domain specificity.
- Comparative evaluation: the first statistically rigorous comparison of GRU, LSTM, and Transformer SANLMs on semantic reasoning over general purpose (e.g., commonsense knowledge) and biomedical NCD KBs.
- New metric: the Meaning-Based Selectional Preference Test (MSPT), which masks the object in an SPO triple and scores the model by how much the embedding-level similarity of its prediction to the gold object phrase exceeds a random-chance baseline, thus capturing semantics beyond exact token matches.
- Qualitative error analysis: a detailed linguistic examination of model errors using semantic frames, roles, and selectional preferences, illuminating domain- and architecture-specific failure modes.

Our results show that attention-based GRU models demonstrated better average generalization according to semantic relatedness measures, while Transformers exhibited superior generalization regarding the average validation loss and accuracy (which were ≈20% vs. ≈53% in the best cases, respectively).

Our MSPT exposes complementary strengths: GRUs generalize semantic relatedness better, achieving a mean similarity of 0.641 with a +0.121 gap (12.1% over a 0.542 random baseline; $p < 10^{-180}$), whereas Transformers reach a mean similarity of 0.551 with a +0.04 gap (4% over a 0.511 random baseline; $p < 10^{-25}$) while being better than GRUs at minimizing and stabilizing loss/accuracy. Although LSTMs show evidence that their generated phrases are not purely random, they still produce senseless sentences that lack contextual relevance to the input subject and object phrases in the NCD domain. Both SANLM architectures operate on commodity hardware (up to 24 GB VRAM gaming GPUs) and effectively learn from noisy, end-to-end knowledge bases. Our approach evaluates semantic coherence in both general-purpose (including commonsense) contexts and non-communicable disease (NCD) domains, offering a computationally efficient alternative to resource-intensive, closed, infeasible, and unaffordable models.

Our findings establish performance baselines for SANLMs in medical knowledge reasoning (both quantitatively and qualitatively), laying a foundation for future research into resource-efficient Artificial Intelligence (AI) systems. Additionally, this framework potentially simplifies the scaling of knowledge-based AI for impactful applications in resource-constrained environments, particularly for public health research, clinical decision support systems, large-scale NCD surveillance, and other public-health tasks.

Section 2 situates our work within the literature; Section 3 provides our rationale for model and data selection; Section 4 reviews theoretical underpinnings; Section 5 describes data and experimental procedures; Section 6 details the proposed pipeline; Section 7 presents and discusses results; Section 8 contains our position statement on noise propagation due to pseudo-labels; and Section 9 concludes and proposes future directions.

2. Related Work

Regarding systems with IE-based semantic reasoning capabilities, we found common traits shared with the seminal work in Chronic Disease Knowledge Graphs (KGs) [6]. The authors proposed a data model to organize and integrate the knowledge extracted from text into graphs (ontologies, in fact), and a set of rules to perform reasoning via first-order predicate logic over a predefined dictionary of entities and relations [7]. More recently, association rule learning was proposed for relation extraction for KG construction [8] and neural network-based graph embedding for entity clustering from EMRs. In Ref. [9], the authors constructed a KG of gene–disease interactions from the literature on co-morbid diseases. They predicted new interactions using embeddings obtained from a tensor decomposition method. The authors of Ref. [10] proposed a KG of drug combinations used to treat cancer, which was built from OpenIE triples filtered using different thesaurus [11,12]. The drug combinations were inferred directly from the co-occurrence of different individual drugs with fixed predicate and disease. The authors created their resource from the conclusions of clinical trial reports and clinical practice guidelines related to antineoplastic agents.

An EMR-based KG was used as part of a feature selection method for a support vector machine to successfully diagnose chronic obstructive pulmonary disease [13]. In recent work, deep learning has been used to predict heart failure risks [14]. The authors used a medical KG to guide the attentional mechanism of a recurrent neural network trained with event sequences extracted from EMRs. Previously, the authors of Ref. [15] also predicted disease risk, but for a broader spectrum of NCDs, and using Convolutional Neural Networks in a KG of EMR events. Medical entity disambiguation is an NLP task aimed at normalizing KG entity nodes, and the authors of Ref. [16] approached this problem as one of classification using the Graph Neural Network. Overall, multiple classical NLP methods have been applied to biomedical KGs, including biomedical KG forecasting from the point of view of link prediction (also known as literature-based discovery) [17].

While relation extraction (RE)-based methods have been widely used to construct KBs [18,19], they often suffer from low recall and limited expressiveness due to their reliance on predefined entity and relation vocabularies [3]. These constraints result in KBs that may not fully capture the richness of biomedical literature. In contrast, OpenIE methods extract relational tuples without the need for predefined schemas, leading to more diverse and comprehensive KBs. For instance, Mesquita et al. [20] demonstrated that OpenIE methods could extract a significantly larger set of relations compared to traditional RE methods. Their study showed that using OpenIE led to improve the overall recall in the extraction of relational tuples, enhancing the coverage and utility of the resulting KBs for downstream tasks. This expansion of relational data provides machine learning models with more diverse and expressive training material, which is crucial for developing robust semantic reasoning capabilities.

3. Rationale for Model and Data Choices

In this study, we employ Standard Attention-Based Neural Language Models (SANLMs), specifically GRU, LSTM networks, and Standard Transformers, for semantic reasoning over knowledge bases constructed via Open Information Extraction (OpenIE). This choice is driven by the need to explore the capabilities of basic attentional mechanisms in object phrase generation within knowledge base reasoning, a task underexplored with such foundational models. Additionally, our use of OpenIE reflects a commitment to resource-efficient, data-independent research, enabling the study of model performance in realistic, noisy knowledge base scenarios, particularly in the medical domain of non-communicable diseases (NCDs) literature.

3.1. Novelty of Basic Attentional Models for Object Phrase Generation

Attentional mechanisms have revolutionized natural language processing, yet their application in their most basic forms—GRUs with attention and Standard Transformers—remains largely unexamined for generating object phrases in knowledge base reasoning. This task requires predicting semantically coherent object phrases (e.g., "insulin" for the subject–predicate pair "diabetes is treated with") within subject–predicate–object (SPO) triples derived from open-vocabulary knowledge bases. While attentional models have been applied to related tasks, such as knowledge graph reasoning [21] and commonsense question answering [22], prior work has primarily focused on complex architectures or alternative model types [5], including Recurrent Neural Networks [23,24], LSTMs [25,26], Latent Feature models [27,28], and Graph-based models [29,30].

Notably, COMET [22] leverages large-scale pre-trained language models, such as GPT variants, to generate commonsense knowledge descriptions (e.g., "a chair is used for sitting"). However, COMET's focus on commonsense knowledge, still constrained by controlled vocabulary (e.g., "it be hot, `HasSubevent`, you turn on fan"), differs significantly from our task of semantic reasoning over (open) medical knowledge bases derived from NCD literature, which involve complex, domain-specific terminology and noisy, heterogeneous data. While GPT and other LLMs demonstrate improved performance metrics, this enhancement may stem from task memorization rather than robust reasoning [22,31,32]. Moreover, these models pose interpretability challenges when fine-tuned for object phrase generation, as their complex pre-trained weights obscure task-specific reasoning mechanisms. Our study addresses this gap by evaluating the intrinsic reasoning capabilities of basic attentional models, providing clearer insights into their potential and limitations in medical knowledge reasoning.

3.2. Comparison to Alternative Approaches

To contextualize our model selection, we compare GRUs with attention and Standard Transformers to alternative methods for knowledge base reasoning, as summarized in Table 1. These methods are less suited for our task due to the noisy, open-vocabulary nature of OpenIE-derived knowledge bases and the requirement to generate natural language object phrases.

Table 1. Comparison of methods for knowledge base reasoning (key limitations for OpenIE-derived knowledge bases are highlighted in bold).

Method Type	**Description**	**Limitations for OpenIE KBs**
Symbolic/Logic-Based (e.g., Description Logics, Prolog)	Use predefined rules or ontologies.	Brittle with noisy data; require extensive manual effort.
Latent feature/embedding (e.g., TransE, DistMult)	Learn vector embeddings only for entities/relations.	Limited in capturing contextual nuances of phrases; designed for fixed vocabularies.
Graph-based (e.g., GNNs, random walks)	Use graph structures for link prediction.	Less suited for generating natural language phrases; focus on individual triples.
LLM-based (e.g., COMET)	Use advanced transformers for knowledge generation (e.g., GPT-based models).	Prone to memorization; resource-intensive; less interpretable.

Symbolic methods struggle with scalability and noise, while embedding-based models like TransE are optimized for structured knowledge graphs. Graph-based methods, such as R-GCNs [29], excel in link prediction but are not tailored for generative tasks involving natural language. COMET, while effective for common sense knowledge, has significant drawbacks: it requires intensive computational resources, tends to memorize rather than reason about meanings, and exhibits reduced interpretability when finetuned for object phrase generation.

In contrast, GRUs with attention and Standard Transformers balance simplicity and power. GRUs excel at modeling sequential dependencies in variable-length SPO phrases, while Transformers' self-attention captures long-range dependencies, ensuring semantic coherence in object phrase generation. These characteristics make them ideal for handling the contextual richness and noise of OpenIE-derived medical knowledge bases.

3.3. Motivation for Using OpenIE

Our use of OpenIE for knowledge base construction is motivated by the goal of creating a resource-efficient, data-independent research framework. Traditional methods for building knowledge graphs or ontologies require significant time, expertise, and financial investment to define schemas and curate relations [5]. OpenIE, by contrast, enables the rapid extraction of relational triples from unstructured NCD literature without predefined schemas, reducing costs and barriers for researchers. This approach simulates real-world scenarios where computational and data resources are limited, testing model robustness in noisy environments and democratizing knowledge-based AI research.

3.4. Noise Mitigation in Knowledge Base Construction

To address potential noise in OpenIE triples, we employed targeted mitigation strategies during the construction of the OIE-GP and OpenNCDKB knowledge bases (see Section 5 for more details). For the OIE-GP KB, we ensured high-quality data by selecting only triples annotated as factually correct from reliable sources, such as ClausIE and MinIE-C, minimizing non-factual or redundant relations. For the OpenNCDKB, we filtered out triples where the subject or object phrase contained only stop words (e.g., "this", "that"), as these are often uninformative or erroneous, reducing the initial 22,776 triples to 18,616 valid ones. This limited preprocessing approach balances data quality with resource efficiency, enabling rapid KB construction from unstructured NCD literature while maintaining semantic coherence for reasoning tasks.

Additionally, we generated 45,032 negative samples using Artificial Semantic Perturbations (like those used in for leveraging our MPTS evaluation) to evaluate our models in distinguishing valid from invalid triples, showing insights into the models' robustness to noise. Regarding this idea, we used object phrase randomization in our MSPT to measure semantic coherence beyond token-based loss and therefore have an idea about noise propagation during learning (see Sections 4.5 and 4.6).

While our noise mitigation strategies are effective for initial evaluations, they represent a minimal preprocessing approach or indirect measurements. Future work could explore advanced techniques, such as triple trustiness estimation [33], which quantifies the reliability of triples based on entity types and descriptions, or canonicalization methods like CESI [34], which reduce redundancy by clustering synonymous phrases. These approaches could further improve model performance in medical knowledge reasoning while maintaining the efficiency of our current methodology.

3.5. Scientific Value and Broader Impact

The scientific value of this work lies in evaluating basic attentional models for object phrase generation in medical knowledge bases, an underexplored area compared to commonsense knowledge tasks. By focusing on GRUs with attention and Standard Transformers, we establish performance baselines that illuminate their reasoning capabilities in handling domain-specific, noisy data. Although the methods we use here are open-domain, this is particularly relevant for NCD literature, where interpretable reasoning is critical for clinical applications. Our use of OpenIE further supports sustainable AI by demonstrating that impactful applications can be developed with cost-effective knowledge resources, with clear performance trade-offs [5]. In addition, our MSPT offers a novel evaluation framework that observes model behavior from the perspective of prediction meaning—beyond mere token matching—and shows promise in assessing semantic coherence. In the future, methods like this may, albeit not entirely, relieve human experts of the burden of validating generated knowledge, thereby facilitating the agile evaluation of knowledge-based systems and expediting their immediate impact in applications for social change or improvement.

4. Theoretical Background

4.1. Open Information Extraction for openKBs

Open Information Extraction (OpenIE) is a paradigm that enables the extraction of relational tuples directly from text without relying on predefined ontologies or relation types. This open-domain and open-vocabulary nature allows OpenIE to capture a wider array of semantic relationships present in biomedical literature, even when classic methods are rule-based [35,36]. The limitations of RE-based methods, such as low recall due to fixed entity and relation sets [2], restrict the expressiveness and contextual richness of the resulting KBs [37]. These constraints can be particularly detrimental when training models for semantic reasoning tasks that require understanding a broad and nuanced range of information.

OpenIE works in such a way that, given the example sentence *"habitat loss is recognized as the driving force in biodiversity loss"*, it generates semantic relations in the form of SPO triples, e.g., {`‘‘habitat loss’’`, `‘‘is recognized as driving’’`, `‘‘biodiversity loss’’`}. In this triple, in the sense of Dependency Grammars [38] (as opposed to Phrase Structure Grammars (the classic approach), where the predicate includes the verb and the object phrase, in Dependency Grammars, the verbal form is assumed to be the center (or highest hierarchy) of the sentence, linking the subject and the object—this latter is a convenient approach in building KBs, and therefore for OpenIE.), ‘‘habitat loss’’ is the subject phrase, ‘‘is recognized as driving’’ is the predicate phrase, and ‘‘biodiversity loss’’ is the object phrase. In semantics, the subject is the thing that performs actions on the object, which is another thing affected by the action expressed in the predicate. Predicates therefore relate things (the subject and the object) in a directed way from the former to the latter.

OpenIE takes a sentence as input and outputs different versions of its SPO structure. Normally these versions are "sub-SPO" structures contained in the same sentence, e.g., {‘‘habitat loss’’, ‘‘is recognized as’’, ‘‘driving biodiversity’’}, and {‘‘habitat loss’’, ‘‘recognized as’’, ‘‘driving’’}. Notice that the last extraction (triplet) may not be factual at all and, depending on the downstream task this kind of output is used for, it may be considered purposeless.

We use the obtained OpenIE triples to build an openKB that organizes knowledge from NCD-related paper abstracts (Section 5.2). In addition, we used already existent extractions to build our OpenIE General Purpose KB (OIE-GP KB) with no specific topics (see Section 5.1). A KB is a special case of a database using a structured schema to store structured and unstructured data. In our case, the structured data constitute the identified elements of semantic triples, i.e., {subject, predicate, object}, while the unstructured part is the open vocabulary text (natural language phrases) of each of these elements.

4.2. Neural Semantic Reasoning Modeling

In this paper, we extend the use of SANLMs originally proposed for Neural Machine Translation (NMT) to learning to infer missing open vocabulary items of semantic structures (SPO triplets). In this section, we show the theoretical background behind the involved methods, which can be seen in Refs. [22,39–41] for more details.

Let $(s, v, o) \in \mathcal{K}$ be a semantic triple, where s, v, o are subject, predicate, and object phrases, respectively, and $\mathcal{K}$ is the training openKB. The KBC task here serves to predict $O = o$ given $U = u = (s, v)$, which gives place to the conditional probability distribution $P(O|U)$ implemented using a neural network model:

$$p(o|u) = f_{nn}(h_o, h_u),$$

where O is the random variable (RV) that takes values on the set of object phrases $\mathcal{O} \ni o$, and U is the RV that takes values on the set of concatenated subject-predicate phrases, i.e., $\mathcal{U} \ni u = s \oplus v$. The neural network $f_{nn}(\cdot, \cdot)$ has learnable parameters $h_o, h_u \in \mathbb{R}^d$, which can be interpreted as phrase embeddings of o and u, respectively. From the point of view of NMT, the probability mass function $p(o|u)$ can be used as a sequence prediction model. In this setting, each word o_i of the target sequence (the object phrase) has a temporal dependency on prior words of the same phrase $o_{i'<i}$, and on the source sequence embedding h_u (the concatenated subject-predicate phrases):

$$p(o_i|o_{i'<i}) = \sigma_d(o_{i-1}, h_i, h_u), \tag{1}$$

where $\sigma_d(\cdot)$ is the decoder activation (a softmax function) that computes the probability of decoding the i-th (current) word of the object phrase from both the current hidden state embedding h_i and the prior state embedding h_{i-1}, as well as from the source embedding. Notice that, in the case of modeling this sequence prediction problem using Recurrent Neural Networks (RNNs), the i index represents time. In NLP, it is simply the position of a word within a sequence of words.

4.3. Encoder–Decoder with Attention-Based Recurrent Neural Networks (Attentional Seq2Seq)

Based on important improvements previously reported [40,42], in this work, we first observe the performance of an encoder–decoder architecture (originally called Recurrent Autoencoder Network) with Attention-Based Recurrent Neural Networks (i.e., an Attentional Seq2Seq model) in neural reasoning tasks. Attentional mechanisms are feature extraction layers of neural network models that improve the expressiveness of inner representations (the encoder's hidden states). This expressiveness encodes the so-called attention weights, which indicate what parts of the input are more important to the prediction task through their inner representations in the network. The result of this inner feature extraction layer is $h_{ui} = c_i \oplus h_i \in \mathbb{R}^{d+d'}$, which is called the attention vector [40]. It is the concatenation, denoted by $\oplus$, of a context vector $c_i \in \mathbb{R}^{d'}$ and the hidden state embedding $h_i \in \mathbb{R}^d$, which are both representations (contextual and temporal, respectively) of o_i.

Endowing an RNN with an attention mechanism makes it necessary to replace Equation (1) with

$$p(o_i|o_{i'<i}, h_{ui}) = \sigma_d(o_{i-1}, h_i, h_{ui}). \tag{2}$$

where the hidden state embedding of o_i is given by

$$h_i = \sigma(h_{i-1}, o_{i-1}, c_i), \tag{3}$$

with $\sigma(\cdot)$ being the hidden state activation of the decoder. There are some usual methods to compute the context vector c_i [40], which encodes the features of the source context in which each word of the output is generated as a target. We used the weighted sum method proposed in Ref. [42]:

$$c_i = \sum_{j=1}^{|u|} \alpha_{ij} h_j, \tag{4}$$

where the attention weights $\alpha_{ij} \in \mathbb{R}$ are computed as

$$\alpha_{ij} = \frac{\exp[g(h_{i-1}, h_j)]}{\sum_{j'} \exp[g(h_{i-1}, h_{j'})]}, \tag{5}$$

and

$$h_j = \sigma_e(h_{i-1}, u_j) \tag{6}$$

is the hidden state of the encoder, so $\sigma_e(\cdot)$ is the corresponding activation. This method, based on an alignment score $g(\cdot, \cdot)$, helps the model learn to pay more attention to a specific pair of embeddings h_j and h_{i-1} when it is expected that their represented words o_{i-1} and u_j co-occur in the input (the subject-predicate concatenation) and in the output (the object) at positions $(i-1, j)$, respectively. In other words, α_{ij} in Equation (5) measures the alignment probability that o_{i-1} will be generated in the next step as a consequence of observing u_j in the current step. Therefore, the weighted sum in Equation (4) measures such a consequential effect $\forall u_j \in u$. In the particular case of our experiments, we used the Attentional Seq2Seq GRU (Gated Recurrent Unit) model proposed by Ref. [40] since GRU-based recurrent models have proven to be effective and better options in computationally constrained environments.

4.4. Transformers

Recurrent models take as input only one element per read. This is the natural way of processing sequences since one element is generated at each timestamp. However, the recently proposed Transformer architecture takes advantage of the fact that the data are already generated and stored, so they can process as many elements of a sequence as possible at once.

As first introduced for Attention-Based RNNs [40,42], we use the encoder–decoder Transformer as an SANLM intended to generate object phrases (output) $o = o_1, \dots, o_n$ given the concatenated subject–predicate phrases (input) $u = s \oplus v = u_1, \dots, u_n$ (notice that, in this case, we have sequences of the same length, n, in the input and in the output). The input Transformer encoder block takes as input the n $d-$dimensional (learnable) word embeddings of each item u_i of the input sequence u in parallel. Therefore, such input to the encoder is a matrix $X \in \mathbb{R}^{n \times d}$ (whose rows are word embeddings x_i) accepted by the ℓ-th attention head of the m-headed multi-head self-attention layer [43]:

$$H_\ell = \langle \sigma(\Lambda), W_v^\top X \rangle \rangle, \tag{7}$$

where $H_\ell \in \mathbb{R}^{n \times (d/m)}$ is the context matrix resulting from the ℓ-th attention head, with $\ell = 1, \dots, m$, and $\sigma(\cdot)$ is the element-wise softmax activation. The attention matrix $\Lambda \in \mathbb{R}^{n \times (d/m)}$ is given by

$$\Lambda = \frac{\langle W_q^\top X, W_k^\top X \rangle}{\sqrt{d/m}}, \tag{8}$$

where $W_q, W_k, W_v \in \mathbb{R}^{d \times (d/m)}$ are fully connected layers with linear activations (simple linear transformation layers), and each entry $\alpha_{ij} \in \mathbb{R}$ of Λ is the attention weight, from u_i to each other u_j.

The multi-head self-attention layer builds its output by concatenating the m context matrices:

$$H_{\oplus \ell} = \bigoplus_{\ell=1}^{m} H_\ell,$$

where then $H_{\oplus \ell} \in \mathbb{R}^{n \times m(d/m)}$ is fed to another linear output fully connected layer, i.e., $H' = W_{\oplus \ell}^\top H_{\oplus \ell} \in \mathbb{R}^{n \times d}$, whose output is in turn fed to the normalization layer given by

$$f_N(z) = \gamma \frac{(z - \mu)}{\varsigma} + \beta,$$

where $f_N(\cdot)$ is applied both to $X + H'$, therefore $H = f_N(X + H')$, and to the output of the block, i.e., $H_b = f_N(H + W_b^\top H) \in \mathbb{R}^{h \times d}$, where $W_b \in \mathbb{R}^{h \times d}$ is an h-dimensional linear output fully connected layer (h is the dimension of the latent space of the encoder–decoder model, i.e., the number of outputs of the encoder; in most cases, $h = n$). The user-defined parameters μ and ς of f_N are the sample-wise mean and variance, i.e., over each input embedding of the layer.

To build the decoder, a second Transformer block is stacked to an input one just after the first normalization layer of the latter (thus, W_b does not operate for the first block). This way, the output of the first decoder block is taken as the query of the second block, whose key and value are the output of the encoder. As in the case of any encoder–decoder configuration, the decoder takes the target sequence as input and output. The Transformer architecture allows the stacking of multiple blocks (layers), which also extends to the encoder and decoder. In this work, we used the $\{1, 2\}-$block Transformer encoder–decoder SANLMs.

4.5. Selectional Preference Analysis for Utterance Plausibility

This section outlines a simplified framework for analyzing subject–predicate–object (SPO) utterances (sentences), focusing on core semantic principles: Frame Semantics, Semantic (or Thematic) roles, and Selectional Preferences. This approach prioritizes objectivity and efficiency for pure linguistic analysis, which is desirable in the analysis of scientific literature.

4.5.1. Frame Semantics and Semantic Roles

The concept of Frame Semantics, introduced by Charles J. Fillmore, posits that word meanings are understood within the context of conceptual structures called "frames". These frames represent typical situations, events, or objects, along with their associated participants and roles.

In this framework, the predicate (verb phrase) serves as the primary activator of a relevant semantic frame. The subject and object are then assigned semantic roles (e.g., Agent, Patient, Theme) based on their relationship to the predicate and the activated frame. This process establishes a structural understanding of the utterance (a sentence in our case), defining the fundamental relationships between its constituents (i.e., the building blocks of a sentence: words, phrases, like noun phrases or verb phrases; or even clauses). The relevance of this process in our manual analysis of a model prediction is that this establishes the basic "who did what to whom" of the sentence, which we compare with the ground truth.

For example, some ground truth sentence can be "The child kicked the ball", from which one can identify its different parts,

- `The` child `kicked` `the` ball;
- `DET` child `PAST`.kick `DET` ball,

where "Kicked" activates the "kicking" frame, from which the following roles are also identified:

- `Agent`: child (the doer, capable of kicking);
- `Patient`: ball (the affected entity, capable of being kicked).

Analysis: in the example, the verb "kicked" brings to mind a frame where a person or thing uses a foot to propel something. The subject "The child" is assigned the role of agent, as it is the entity performing the action. The object "the ball" is assigned the role of patient, because it is the entity that is being propelled (so being affected by the action). Semantic plausibility arises when the assigned thematic roles align with our real-world knowledge and expectations about how events typically unfold. In the particular case of this example, the Agent role implies a capacity for action (child: the doer, is capable of kicking), so an animate entity is typically expected (a human). Therefore, the Patient role implies a capacity for receiving action (ball: the affected entity, is capable of being kicked), so an inanimate entity is expected (an object). In this case, the expectations are filled and the sentence is semantically plausible because children typically kick balls.

In cases where semantic plausibility is not verified, expectations based on real-world knowledge are not filled. For example, in "The table analyzed the data",

- `Agent`: table (inanimate, incapable of analysis);
- `Patient`: data (abstract).

Analysis: The thematic role assignment violates expectations based on real-world knowledge. A table (inanimate $\notin$ Agent) cannot perform the action of analyzing. This violates our understanding of the capabilities of a table. The thematic role of the Agent requires certain capabilities that a table does not possess. Therefore, the sentence fails in terms of semantic plausibility.

4.5.2. Selectional Preferences

Selectional preferences are constraints on the types of arguments that a verb or other predicate can take. They describe the semantic restrictions that a word imposes on the words that can appear with it. This linguistic analysis framework focuses on the semantic compatibility between a predicate and its arguments. The relationship among the different linguistic theories described in this section is that selectional preferences can be seen as linguistic manifestations of the thematic roles and constraints defined by semantic frames. Therefore, from an end-to-end perspective, frames identified by Frame Semantics analysis provide the basis for preferences. As Resnik explains,

> "Although *burgundy* can be interpreted as either a color or a beverage, only the latter sense is available in the context of *Mary drank burgundy*, because the verb *drink* specifies the selection restriction `[liquid]` for its direct objects". [44].

Testing utterances (sentences in our case) for the validation of selectional preferences conveys an analysis where we verify the semantic compatibility between the arguments and their assigned semantic roles [45]. For instance, in the sentence *"The dog ate the bone"*.

- `Frame`: Eating
- `Roles`:
  - `Agent`: dog (animate, capable of eating)
  - `Patient`: bone (edible, capable of being eaten)
- `Constraints`: dog $\in$ animate, and bone $\in$ edible (both constraints are satisfied).

Analysis: The verb *"eat"* imposes preferences for an [animate] Agent and an [edible] Patient. The nontransitivity property of the verb allows for semantic plausibility even when Patient is absent. Therefore, the sentence is semantically valid.

Violations of the constraints means that preferences are not met, therefore indicating potential semantic anomalies. For example, in the sentence *"The rock ate the bone"*, we have

- `Frame`: Eating
- `Roles`:
  - `Agent`: rock (inanimate)
  - `Patient`: bone (edible)
- `Constraints`: rock $\notin$ `[animate]` (conflicts with the required `[animate]` Agent).

Analysis: The verb *"eat"* imposes preferences for an `[animate]` Agent and an `[edible]` Patient, and these constraints are partially meet by the Patient (bone). However, the Agent constraint is violated because a rock is inanimate, which conflicts with the required [animate] Agent. Therefore, this sentence is semantically anomalous.

4.6. Meaning-Based Selectional Preference Test (MSPT)

When evaluating the reasoning capabilities of language models, traditional performance metrics that rely on exact token matching or token-based statistics may not effectively capture the nuances of natural language semantics and built meanings. This is particularly true given the open vocabulary inherent in natural language and the semantic variability arising from logical inference, synonymy and context. For instance, two phrases can convey the same meaning using different words or structures, rendering exact matching insufficient for assessing semantic reasoning. Moreover, reliance on exact token matching—which drives most current evaluation methods—can lead to the appearance of false positive "hallucinations" [46].

To address these challenges, we introduce a probe that hides the object phrase in withheld subject–predicate–object triples, asks the model to supply that object, and then scores the answer by how much its phrase-embedding similarity to the gold object exceeds a random-choice baseline—so higher MSPT gaps reflect stronger, genuinely semantic selectional preferences rather than lucky token matches, i.e., Meaning-Based Selectional Preference Test (MSPT). This method is designed to measure reasoning quality by assessing semantic relatedness (particularly, meaning similarity) [47,48], specifically focusing on the selectional preferences between subjects, verbs, and objects that drive semantic interpretation [44,49].

From the previous Section 4.5, selectional preferences are semantic constraints that certain verbs impose on their arguments, determining which types of nouns are appropriate as their subjects or objects. For example, the intransitive verb *eat* typically selects for objects that are edible; contrary to the transitive verb *attend*, which mostly selects for a place as object, e.g., *to the school*. Capturing these preferences is crucial for understanding and generating coherent and contextually appropriate language.

In our experiments, we introduce the notion selectional preferences considering phrase meaning, by means of the word embeddings that in turn build phrase embeddings. Let $x_a, x_b \in \mathbb{R}^d$ be the phrase embeddings of two phrases S_a and S_b, respectively. These embeddings are obtained by summing the word embeddings of the words in each phrase:

$$x_a = \sum_{w \in S_a} x_w,$$

$$x_b = \sum_{w \in S_b} x_w,$$

where $x_w \in \mathbb{R}^d$ is the word embedding of word $w \in S_{(\cdot)}$. The phrase similarity between S_a and S_b is measured using the cosine similarity:

$$\cos(\theta) = \frac{\langle x_a, x_b \rangle}{\|x_a\| \cdot \|x_b\|},$$

where $\langle x_a, x_b \rangle$ denotes the dot product of x_a and x_b, and $\|x_a\|$ is the Euclidean norm of x_a. The cosine similarity $\cos(\theta) \in [-1, 1]$ quantifies the degree of semantic similarity between the two phrase embeddings, with values closer to 1 indicating higher similarity.

To evaluate the average semantic similarity across a set of m phrase pairs, we compute the mean cosine similarity:

$$\mu_{\text{sts}} = \frac{1}{m} \sum_{k=1}^{m} \frac{\langle x_a^{(k)}, x_b^{(k)} \rangle}{\|x_a^{(k)}\| \cdot \|x_b^{(k)}\|}.$$

This mean value μ_{sts} reflects the average semantic relatedness between the meanings of the model's predicted phrases $x_a^{(k)}$ and the meanings of the reference/baseline phrases $x_b^{(k)}$, with $k = 1, \ldots, m$.

To assess the reasoning capabilities of the language model, we employ the Student's *t*-test to compare the distribution of STS measurements between the predicted object phrases and shuffled ground truth object phrases. The shuffled object phrases serve as a random baseline, simulating a scenario where the correspondence between the subject–predicate pair and the object is disrupted. This randomization effectively perturbs the selectional preferences that are pivotal for semantic coherence.

By disrupting these preferences, we assess whether the model can distinguish between semantically coherent and incoherent phrase tuples. Specifically, we analyze whether the model's predictions are significantly more semantically similar to the ground truth phrases than to the randomized ones. For example, from a previous section, the word *burgundy* can refer to either a color or a beverage. However, in the frame induced by the verb drink, only the beverage sense is appropriate for *burgundy* due to the selectional preference of the verb *drink*, which specifies a semantic constraint for its object to be of the type `[liquid]`. Our random baseline becomes the dataset agnostic of these constraints.

Violating selectional restrictions creating semantically incoherent subject–predicate⇒ object triples enables us to quantitatively evaluate how effectively the language model understands and applies selectional preferences during reasoning. In addition, the use of word embeddings for meaning representation allows us to consider the inaccuracy of natural language.

The hypothesis testing framework is as follows (see Figure 1):

- Null hypothesis (H_0): There is no significant difference in semantic similarity between the model's predictions and the ground truth phrases versus the shuffled phrases. Under H_0, the model's ability to predict/generate object phrases is equivalent to random chance, indicating a lack of semantic reasoning based on meaning and selectional preferences.
- Alternative hypothesis (H_1): The model's predicted/generated object phrases are significantly more semantically similar to the ground truth phrases than to the shuffled phrases. This suggests that the model effectively captures selectional preferences for meanings and demonstrates genuine semantic reasoning.

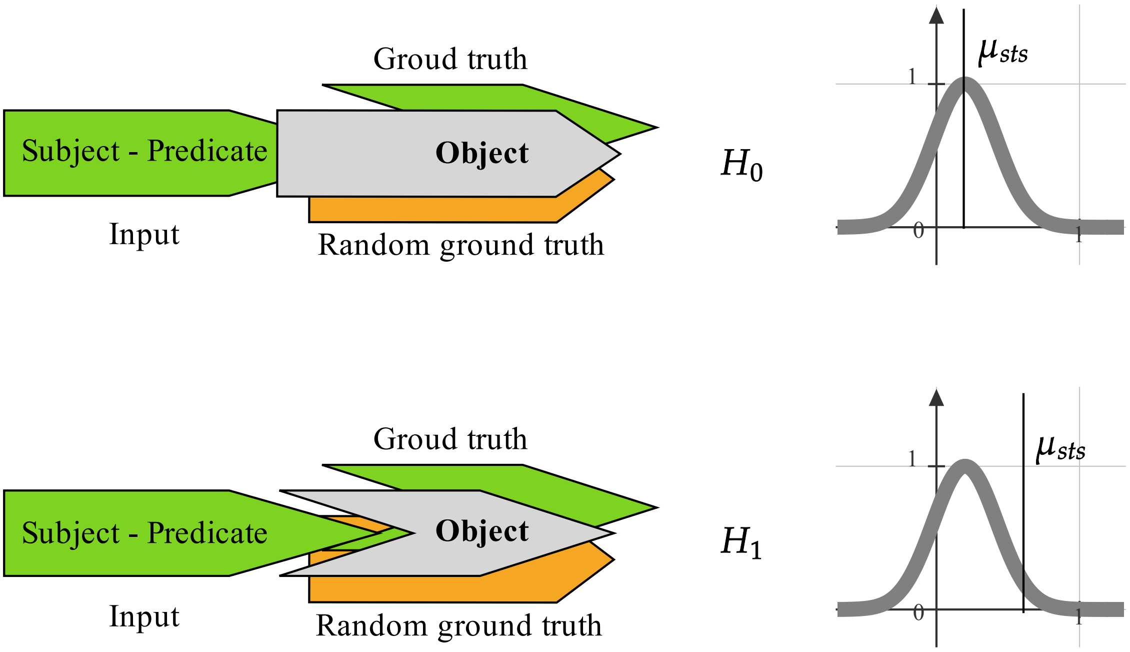


Figure 1. Depiction of our MSPT method.

The hypothesis testing framework is as follows (see Figure 1):

By performing the t-test on the STS measurements, we determine whether the observed mean similarity μ_{sts} between the model's predictions and the ground truth is statistically greater than that with the randomized baseline. A significant result in favor of H_1 indicates that the model's reasoning is not merely a product of surface-level associations but reflects an understanding of deeper semantic relationships dictated by selectional preferences.

This evaluation method emphasizes the model's ability to generate meanings that are contextually and semantically appropriate, moving beyond mere token-level accuracy. It provides a robust framework for assessing semantic reasoning in language models, especially in open-vocabulary settings where lexical variability and synonymy are prevalent.

By incorporating selectional preferences and STS into our evaluation, we align the assessment more closely with human language understanding, where context and semantic constraints play crucial roles. This approach enhances the interpretability of evaluation results and offers deeper insights into the semantic capabilities of language models.

5. Data

5.1. The OIE-GP Knowledge Base

5.1.1. Dataset Description

In this work, we created a dataset called OpenIE General Purpose KB (OIE-GP) using manually annotated and artificially annotated OpenIE extractions. The main criterion for selecting the sources of the extractions was that they had some human annotation, either for identifying the elements of the structure or for their factual validity. We considered factual validity as an important criterion because it is important to train SANLMs to reason with factual validity.

ClausIE: This dataset was used to generate a large number of triples that were manually annotated according to their factual validity [50] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/ambiverse-nlu/clausie (accessed on 2 June 2025). The resulting dataset provides annotations indicating whether the triples are either too general or senseless (correct/incorrect), which recent datasets have adopted in some way. For example, it is common to find negative samples (incorrect triples) like, e.g., `{‘‘he’’; ‘‘states’’; ‘‘such thing’’}`, `{‘‘he’’; ‘‘states’’; ‘‘he’’}`. From this dataset, we took the 3374 OpenIE triples annotated as correct (or positive) samples.

MinIE-C (MinIE-C is the less strict version of MinIE system, and we selected it as it is not restricted in the length of the slots of the triples: https://github.com/uma-pi1/minie (accessed on 2 June 2025)) provided artificially annotated triples that are especially useful to our purposes because they are as natural as possible in the sense of open vocabulary [51]. These are generated by the ClausIE algorithm (as part of MinIE) and are annotated as positive/negative according to the same criterion used by ClausIE. From this dataset, we took the 33216 OpenIE triples annotated as positive samples.

CaRB is a dataset of triples whose structure has been manually annotated (supervised) with n-ary relations [52]. From this dataset, we took the 2235 triples annotated as positive samples.

WiRe57 contains supervised extractions along with anaphora resolution [53]. The 341 hand-made extractions of the dataset are 100% useful as positive samples because they include anaphora resolution.

5.1.2. Noise Mitigation Strategy

Annotation-based selection: The datasets publicly available already include annotations as factually correct or factually incorrect. Only triples annotated as factually correct from sources like ClausIE (3374 triples), MinIE-C (33,216 triples), CaRB (2235 triples), and WiRe57 (341 triples) were selected. This ensures semantic validity and reduces non-factual or senseless triples.

5.2. The OpenNCDKB

5.2.1. Dataset Description

A noncommunicable disease (NCD) is a medical condition or disease that is considered to be non-infectious. NCDs can refer to chronic diseases, which last long periods of time and progress slowly. We created a dataset of scientific paper abstracts related to nine different NCD domains: breast cancer, lung cancer, prostate cancer, colorectal cancer, gastric cancer, cardiovascular disease, chronic respiratory diseases, type 1 diabetes mellitus, and type 2 diabetes mellitus. These are the most prevalent worldwide NCDs, according to the World Health Organization [54].

The domains from which we built our dataset of scientific paper abstracts were for nine different NCDs:

We used the names of the diseases as search terms to retrieve the k∼150 most relevant abstracts from the PubMed (National Library of Medicine, National Center for Biotechnology Information (NCBI): https://pubmed.ncbi.nlm.nih.gov (accessed on 2 June 2025)) (see Table 2). The resulting set of abstracts constituted our NCD dataset. To generate our Open Vocabulary Chronic Disease Knowledge Base (OpenNCDKB), we retrieved a total of 1200 article abstracts that correspond to the NCD-related domains mentioned above.

The domains from which we built our dataset of scientific paper abstracts were for nine different NCDs:

Table 2. Different target domains included in the OpenNCDKB and the number of abstracts retrieved from PubMed.

Disease	Subtype	Abstracts (k)
Cancer	Breast Cancer	100
	Lung Cancer	100
	Prostate Cancer	100
	Colorectal Cancer	100
	Gastric Cancer	100
Cardiovascular Disease		150
Chronic Respiratory Diseases		150
Type 1 Diabetes Mellitus		200
Type 2 Diabetes Mellitus		200
Total		1200

To obtain OpenIE triples (and therefore, an open vocabulary KB), we used the CoreNLP and OpenIE-5 libraries [55,56]. First, we took each abstract from the NCD dataset and split it into sentences using the coreNLP library. Afterwards, we took each sentence and extracted the corresponding OpenIE triples using the OpenIE-5 library. By doing so, we obtained a total of 22,776 triples.

5.2.2. Noise Mitigation Strategy

The 22,776 triples obtained from OpenIE were filtered to remove those containing only stop words in the subject or object phrases. After this preprocessing, we were left with 18,616 triples that were considered valid for our purposes.

In addition to the valid triples, we also generated semantically incorrect negative samples (triples). These were generated using the same methods for artificial semantic perturbations and were preprocessed in the same way as the positive ones, giving 45,032 semantically invalid triples. These negative samples were used for quantitatively evaluating selectional preferences (see Section 4.6), and can be used for teaching factual validity to the models in future work, which resemble the plausibility score included in the ConceptNet Common Sense Knowledge Base [57].

5.3. Our Semantic Reasoning Tasks

To perform training, we first used a compact version of the ConceptNet KG converted into a KB (the ConceptNet Common Sense Knowledge Base [57]) (https://home.ttic.edu/~kgimpel/commonsense.html (accessed on 2 June 2025)) consisting of 600k triples. We also used our OIE-GP Knowledge Base (Section 5.1) containing 39,166 triples, i.e., only the (semantically) positive samples of the whole data described in Section 5.

From these KBs, we constructed mixed KBs that include the source task vocabulary and include, as much as possible, the missing vocabulary needed to validate the model on the target task related to NCDs. In this way, we obtained our source KBs:

1. **OpenNCDKB**. We split the 18.6k triples of the OpenNCDKB into 70% (13.03k) for training and 30% for testing (5.58k). We included the OpenNCDKB here because, in the context of the source and target task, we simply split the whole KB into train, test, and validation data. Validation data was considered the target task in this case.
2. **ConceptNet+NCD**. By merging the triples collected from the ConceptNet Knowledge Graph and those from OpenNCDKB, we obtained our ConceptNet+NCD KB containing 429.12k training triples and 183.91k test triples.
3. **OIE-GP+NCD**. We obtained our OIE-GP+NCD KB by merging 39.17k OIE-GP and 13.03k NCD triples to obtain a total of 52.20k triples. These were split into 70% (36.54k) for training and 30% (15.66k) for testing.
4. **ConceptNet+OIE-GP+NCD**. We obtained this large KB that included ConceptNet, general purpose OpenIE (the OIE-GP KB), and NCD OpenIE triples (the OpenNCDKB) by taking the union between ConceptNet+NCD and OIE-GP+NCD. This source training task constitutes 600k + 39.17k + 13.03k = 652.20K total triples. These were split into 70% (456.54k) for training and 30% (195.66k) for testing.

All the mentioned quantities consider that we filtered out the triples whose subject or object phrases were only stopwords. The utilization of OpenIE extractions in constructing the OIE-GP and OpenNCDKB Knowledge Bases ensures that our dataset captures a broad spectrum of semantic relations present in the biomedical literature. This comprehensive approach aligns with our objective to overcome the limitations of RE-based KBs and provides a robust foundation for training our semantic reasoning models.

6. Methodology

DL and NLP researchers recently tested Transformer-based [21,22,41] and recurrent neural network-based [57] SANLMs in general purpose CSKR tasks where open vocabulary is considered. In sight of this progress, our research uses a methodological approach that can be especially useful for the semantic analysis of documents dealing with arbitrary but specialized topics, such as open domain scientific literature.

Building upon the advantages of OpenIE, our methodology integrates these methods to construct the training KBs for our semantic reasoning models. By extracting relational tuples from biomedical literature using Stanford CoreNLP [55], we generated the OpenNCDKB that encompasses a wider range of semantic relations from a dataset of paper abstracts retrieved from PubMed. We also integrate in our experiments already extracted relations from general purpose texts and with different state-of-the-art OpenIE methods described in Section 5.1.

By using our KBs, we trained Transformer and recurrent encoder–decoder SANLMs to address combinations of two similar semantic reasoning tasks that involve general purpose and open vocabulary: the ConceptNet common sense KB, and the OIE-GP, an openKB we built from multiple sources that used OpenIE extractions. The train, test, and validation KBs were enriched with specialized domains, i.e., chronic disease literature abstracts induced via the OpenNCDKB.

Although this open vocabulary approach can add complexity to modeling semantic relationships (and therefore to the learning problem), it also adds expressiveness to the resulting KBs. Such expressiveness can improve the contextual information in semantic structures and thus allow the SANLMs to discriminate useful patterns from those that are not for the semantic reasoning task [58]. We use Standard Transformer and Attention-Based GRU models because they have been instrumental in the advancement of modern language modeling, still demonstrating state-of-the-art performance in diverse NLP tasks. The Transformer architecture, with its self-attention mechanism and parallelization capabilities, has significantly improved the ability to capture long-range dependencies and contextual relationships. Similarly, attention-based GRUs enhance sequential modeling by dynamically weighting input–output relevant information, making them effective for capturing semantic nuances. Given their strengths in contextual reasoning and semantic representation, we believe these models are well-suited for addressing fundamental reasoning questions at the semantic level. In addition to these models, we include attention-based LSTM networks in order to have baseline comparison.

To show the effectiveness of our approach from multiple points of view, we evaluated and analyzed the results for both the Transformer and the recurrent SANLMs in terms of

1. **Performance metrics** on train and test data. In particular, accuracy considers the reasoning quality as an exact prediction of the model with respect to the ground truth tokens. Cross entropy quantifies the degree of information dissimilarity between the probability distribution predicted by the model and the ground truth one. These metrics are effective at learning time because the inner representations of the SANLM acquire knowledge on what are the specific words (for the object phrase in our case) that probably should be next to the input ones (the subject and predicate). This helps to select models for subsequent evaluations and to establish baseline models.
2. **Attention matrix visualization.** In our specific context dealing with semantic reasoning tasks, visualizing the attention matrices using heatmaps would help in understanding how well the models are capturing the semantic nuances and relationships in the data. This provides interesting insights when dealing with complex, domain-specific data like medical literature, where understanding the context and relationships between concepts is crucial. This helps to further select models for subsequent evaluations.
3. **Meaning-based selectional preference test (MSPT).** To evaluate the semantic reasoning quality of model predictions, we conducted hypothesis testing using STS and selectional preference perturbations on test and validation datasets. This approach measures the semantic relatedness between model-generated object phrases, ground truth references, and a random baseline (Section 4.6). Specifically, we designed a three-way comparison framework to assess
   (a) The similarity distribution between predicted and ground truth object phrases;
   (b) The separation of this distribution from that of shuffled ground truth phrases (random baseline) which simulates misalignment of selectional preferences.

   STS scores were calculated to quantify meaning alignment, with a focus on whether predictions significantly diverged from misaligned selectional preferences and rather converged toward gold standard references. Statistical analysis was applied to determine the significance of differences between these distributions.
4. **Manual inspection of the inferences** for the test and validation data. To complement quantitative evaluations, we conducted a manual inspection of model inferences across test and validation datasets. This qualitative analysis framework aims to
   (a) Identify semantic patterns captured by SANLMs through their generated outputs;
   (b) Assess the consistency of these patterns across test and validation model predictions.

   By systematically comparing inferred outputs against ground truth references, we evaluated how robustly SANLMs encode semantic regularities (e.g., contextual relationships, lexical coherence) and whether these generalize beyond training data.

From the perspective of using the ConceptNet knowledge base (KB) as training data, it is important to note that, while this KB has been largely compiled manually, it was not originally designed for the direct training of models to generate triplets. Regarding IOE-GP and OpenNCDKB, these KBs are built using extractions of already trained supervised and unsupervised methods, neither of which were designed for direct training of models to generate triplets. Therefore, the training semantic reasoning tasks to address by our SANLMs represent a distantly supervised learning problem for predicting/generating object phrases, given the subject and predicate phrases.

7. Results and Discussion

7.1. Experimental Setup

For our experiments, we used previously validated Transformer and attention-based GRU model hyperparameters to evaluate their performance on our multiple semantic reasoning tasks, i.e., predicting/generating object phrases, given the subject and predicate phrases (see Table 3). Due to the fact that our KBs are smaller than the datasets used for Neural Machine Translation (NMT) by the authors of [39,40] to introduce these models, we decided to use the smallest architectures they reported.

In the case of the standard Transformer model [39], we used a source and target sequence lengths of $\max|u| = \max|o| = 30$, a model dimension of $d = 512$ (the input and positional word embeddings), an output Feed Forward Layer (FFL) dimension of $h = 2048$ (denoted as d_{ff} in the original paper), a number of attention heads of $m = 8$, attention key and value embedding dimensions of $d/m = 64$, and a number of transformer blocks of $N = 2$. In addition, we considered the possibility that such a "small" model is still too big for our KBs (the largest one has ∼652k triples) compared to the 4.5 million sentence pairs this model consumed in the original paper for NMT tasks. Therefore, we also included an alternative version of the base model using only one Transformer block ($N = 1$) in both the encoder and decoder (in the case of the decoder, a single Transformer block refers to two self-attention layers ($N = 1$), whereas $N = 2$ refers to three of these layers (i.e., the decoder's number of blocks is $N + 1$ with respect to the encoder).), while keeping all other hyperparameters of the $N = 2$ model.

Table 3. Hyperparameter configurations for model architectures (key contrasting architectural configurations are highlighted in bold).

	Recurrent		Transformer	
Parameter	**GRU Wide**	**GRU Squared**	**N = 1**	**N = 2**
Seq. Len. (u/o)	***10/10***	***30/30***	30/30	30/30
Embedding Dim. (d)	***256***	***1024***	512	512
Hidden Units (m)	***2048***	***1024***	–	–
FFL Dim. (h)	–	–	2048	2048
Heads (m)	–	–	8	8
Blocks (N)	–	–	***1***	***2***
Key/Value Dim.	–	–	64	64
Training KBs	OIE-GP+NCD, CN+OIE-GP+NCD		All Four Mixed KBs	
Number of Models	4 Baseline Models		8 Total Models	

In the case of the attention-based GRU model, we adopted the hyperparameters specified in the corresponding original proposals [40,42] to build baseline recurrent models. In addition, we considered previous work where the additional architecture hyperparameters of Transformers and recurrent models were compared in their performance [59]. Namely, we included two contrasting recurrent architectures (for both encoder and decoder): the first one was labeled as GRU-W (GRU Wide): sequence lengths of $\max|u| = 10$ and $\max|o| = 30$, number of hidden states (units) $m = 2048$, each with an embedding dimension of $d = 256$, and training batch size of 128. The second one was labeled GRU-S (GRU Squared): sequence lengths of $\max|u| = 10$ and $\max|o| = 30$, number of hidden states (units) $m = 1024$, each with an embedding dimension of $d = 1024$, and training batch size of 32.

The models were trained using different KBs constructed from different sources to select the semantic reasoning task and the model that best generalizes the OpenNCDKB validation set. Using the four source KBs and the OpenNCDKB, we obtained four mixed KBs used for training the models (see Section 5.3): ConceptNet+NCD (CN+NCD, for short), OIE-GP+NCD, Concepnet+IOE-GP+NCD (CN+OIE-GP+NCD, for short). Using each of these KBs, we trained eight Transformer models, four models with $N = 1$ (i.e., Transformer 1) and four models with $N = 2$ (i.e., Transformer 2). In the case of the baseline recurrent models, GRU-W and GRU-S, these were trained with the OIE-GP+NCD and CN+OIE-GP+NCD KBs separately. Thus, we trained four baseline models across different KBs, contrasting architectural dimensionalities through two configurations: GRU-W (wide: large hidden units with smaller embeddings) and GRU-S (squared: balanced hidden/embedding dimensions).

We compared and analyzed the test semantic reasoning tasks using performance metrics (i.e., sparse categorical cross-entropy) of the eight Transformer models and the four GRU models (training and test). We set 40 epochs max. to train and test the models with each KB, but with patience=10 for a minimum improvement of $\Delta = 0.005$.

We also performed an STS-based hypothesis test on all resulting test and validation data (see Section 4.6). The overall outcome of this STS-based hypothesis test was to verify whether the STS measurements with respect to the true object phrases and with respect to random baselines come from different distributions; that is, to decide whether the null hypothesis, i.e.,

Hypothesis 0 (H_0)**.** *semantic similarity measurements with respect to true object phrases and with respect to random baselines come from the same distribution, can be rejected with confidence and whether this holds for both the test and validation datasets.*

The neural word embeddings we used for Meaning-Based Selectional Preference Test (MSPT) were trained on the Wikipedia corpus to obtain good coverage of the set difference between the vocabulary of the test and validation data (PubMed paper abstracts contain a simpler vocabulary than the paper itself, while Wikipedia contains a relatively technical vocabulary). The word embedding method used was FastText, which has shown better performance in representing short texts [60]. The phrase embedding method used to represent object phrases was a simple embedding summation; this was because, at the phrase level, even functional words (e.g., prepositions, copulative and auxiliary verbs) can change the meaning of the represented linguistic sample [48].

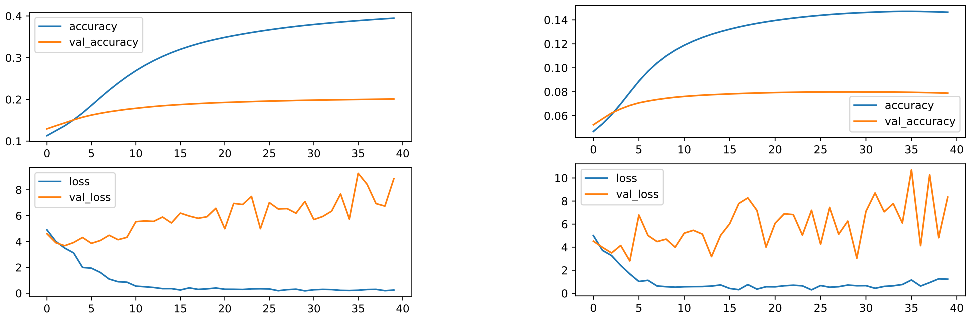
Additionally, we investigate how the demands of semantic reasoning tasks differ from those of neural machine translation (NMT), where GRU hyperparameters have been extensively validated. To this end, we conducted a sensitivity analysis for attention-based GRU and LSTM models trained on our largest combined KB (CN+OIE-GP+NCD) to provide a broader performance evaluation of these standard attentional models, which are less explored in reasoning tasks compared to Transformers. The analysis evaluates the impact of embedding dimensionality, number of hidden units (a.k.a. steps, or states), number of layers, and dropout probability (regularization) on accuracy and loss. To this end, we conducted two sensitivity studies—one for the attention-based GRU (311 trials) and the other for the attention-based LSTM (168 trials). All models were trained on the same KB, ran on four RTX-4090 GPUs, and were tuned with Bayesian optimization; a random forest surrogate model supplied the feature-importance and the Spearman correlations (for all our hyperparameter search analysis, we used the weights and biases (https://shorturl.at/SHw6Y (accessed on 2 June 2025)) Python API).

At the end, we manually inspected and discussed the natural language predictions of the best models using both test and validation data. To do this, we randomly selected five input samples from the test subject–predicate phrases and fed them to the SANLMs that exhibited the highest confidence during our MSPT. We analyzed the predicted object phrases to explore their meaning and the semantic regularities they demonstrated. We then repeated this inspection using validation data from the OpenNCDKB, randomly selecting five subject–predicate inputs not seen during training. This allowed us to verify whether the semantic regularities identified in the test predictions were reproduced in the validation predictions.

7.2. Results

7.2.1. Training- and Validation-Loss Profiles of Attention-Based GRU and Transformer Models

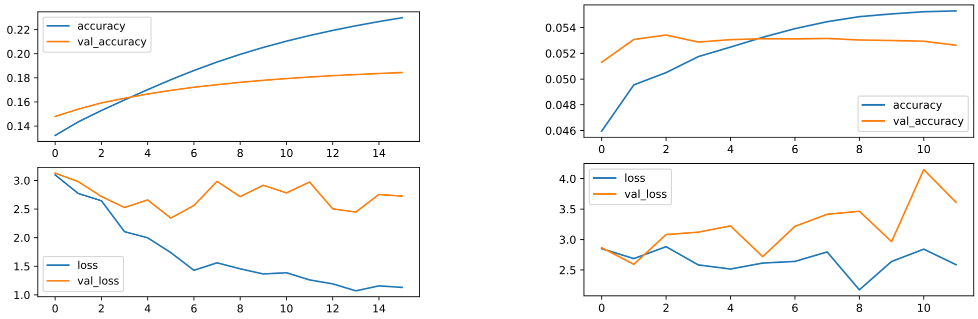
In Figure 2, we start by showing the progress of the performance metrics of the two baseline GRU models (GRU-W and GRU-S) during 40 epochs. This time, both models were trained and validated on the OIE-GP+NCD KB. Figure 2a shows the accuracy and the loss for the GRU-W. Notice that the difference between its train and validation accuracy (`accuracy` = 39% and `val_accuracy` = 20%, respectively) is relatively large, stable, and continues to diverge slowly. Something different occurred for the loss function (`loss` = 0.28 bits and `val_loss` = 8.25 bits). The divergence is even more prominent while the validation loss (`val_loss` in the figure) appears much more unstable even though the training loss settles relatively early. This can be the manifestation of overfitting, although this model showed to be relatively stable through the fourteen epochs allowed by the patience hyperparameter. Notice in Figure 2b that a very similar pattern can be seen in the GRU-S model, but it shows a more unstable loss when making predictions (even reaching 10 bits) and with much less accuracy (`val_accuracy` = 7.8%). Also note that both models are trained on a small KB (36.54k tuples), which we believe is the main cause of this divergent behavior.



(**a**) (**b**)

Figure 2. Training and validation (`val_`) accuracies (upper plots), and training and validation (`val_`) losses (lower plots) for the GRU models trained with the OIE-GP+NCD KB: (**a**) GRU-W; seq. len = 10, batch size = 128, embedding dim. = 256, units = 2048, (**b**) GRU-S; seq. len = 30, batch size = 32, embedding dim. = 1024, units = 1024. For this dataset, we set 40 epochs maximum.

In Figure 3, we have the GRU-W and GRU-S models but now trained on the CN+OIE-GP+NCD KB, which is a larger KB (456.54k train triples). In general, due to the size of this KB, the first improvement we see is that the training and validation curves are much less divergent (both in terms of accuracy and loss). In addition, the instability of the models when making predictions was significantly reduced. In particular, the GRU-W model decreased its validation loss to 2.72 bits with the CN+OIE-GP+NCD KB (Figure 3a), compared to 6.19 bits with the OIE-GP+NCD KB in epoch number 15 (allowed by the patience hyperparameter this time). Furthermore, the difference between the training and validation losses is considerably smaller with the CN+OIE-GP+NCD KB, i.e., 2.72 − 1.13 = 1.59 bits (versus 6.19 − 0.25 = 5.94 bits with the OIE-GP+NCD KB). The training and validation accuracies also show much less difference, and overall, we see that the effect of the training KB size turns out to be a significant reduction in model overfitting in this semantic reasoning task.



(**a**) (**b**)

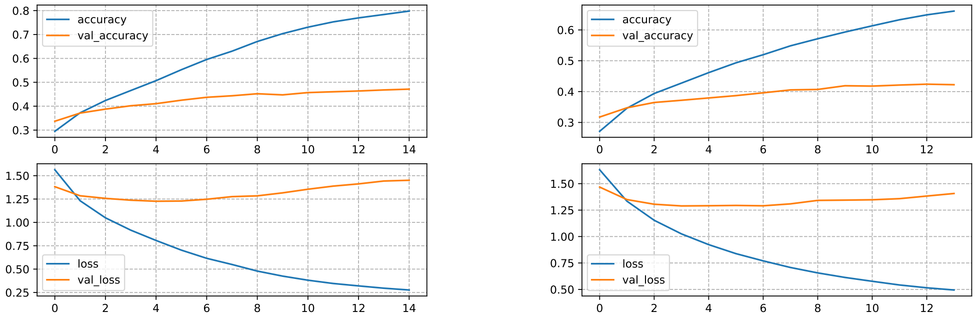
Figure 3. Training and validation (`val_`) accuracies (upper plots), and training and validation (`val_`) losses (lower plots) for the GRU models trained with the ConceptNet+OIE-GP+NCD KB: (**a**) GRU-W; seq. len = 10, batch size = 128, embedding dim. = 256, units = 2048, (**b**) GRU-S; seq. len = 30, batch size = 32, embedding dim. = 1024, units = 1024. For this KB, we set 100 epochs max., but with patience = 10 for a minimum improvement of Δ = 0.005.

Although the training and validation losses of the GRU-S model resulted in less divergent tendencies (Figure 3b), see that these actually increase as the model is exposed to more epochs. See also that the precision of the corresponding validation loss is small and practically constant, without improvement. We therefore observe that architectural differences between the recurrent models (wide and squared) drive their contrasting behaviors. While GRU-W is characterized by a large number of hidden units (2048) of lower dimensionality (256), the GRU-S model is characterized by half of these hidden units (1024), but 3 times higher dimensionality (1024). Furthermore, GRU-W must predict sequences of 10 words, while GRU-S must predict sequences of 30 words. From these experiments with contrasting attention-based recurrent architectures, we observe that increasing the number of hidden units and reducing their dimensionality is what allows an attention-based recurrent model to improve its generalization in this semantic reasoning task.

In Figure 4, we show the progress of the performance metrics of the standard Transformer models trained and validated from scratch only on the OpenNCDKB (13k + 5.6k triples, respectively). Figure 4a shows the accuracy and the loss for the one-block Transformer model (Transformer 1). Notice that the difference between train and validation accuracy (`accuracy` = 80% and `val_accuracy` = 48%, respectively) is relatively large. What occurred for the loss function with the train and validation losses is a bit different because 1.03 bits is a normal separation for two relatively small values (`loss` = 1.28 bits and `val_loss` = 0.25 bits). See that, in the figure, both values are actually near from 0.85 bits. This can be seen as a stable generalization pattern with a relatively small dataset in forty epochs allowed by the patience hyperparameter.

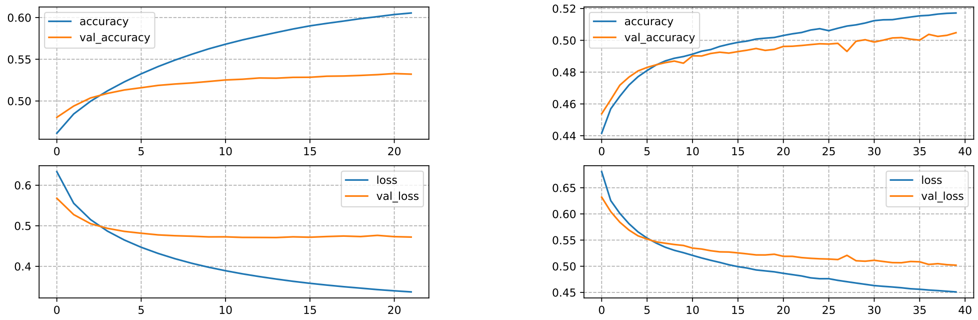
Similarly to the one-block model, the accuracy and the loss (Figure 4b) for the two-block Transformer (Transformer 2) shows a divergence between the train and validation accuracies but with values reduced by 10% with respect to the Transformer 1. This model stopped to improve two epochs earlier than the one-block model, also with their losses around 0.85 bits and with less than 1 bit of difference.

The models trained with the CN+NCD KB were provided with much more data (429.12k training triples) than the models only using the OpenNCDKB (13k training triples). The accuracy and loss for these models are shown in Figure 5. Figure 5a shows a clear improvement in the validation loss, reaching a minimum near 0.48 bits. On the other hand, the training loss was about 0.13 bits apart from the validation loss, which is less than the same comparison made for the two-block Transformer trained only with the OpenNCDKB. Regarding accuracy, the maximum on validation data was about 53%, showing similar train and validation curves with respect to OpenNCDKB but much less divergent, around 5% (compared to 30%) in 20 training epochs (60% more training steps).



(**a**) (**b**)

Figure 4. Training and validation (`val_`) accuracies (upper plots), and training and validation (`val_`) losses (lower plots) for the Transformer trained with the OpenNCDKB: (**a**) N = 1, (**b**) N = 2.

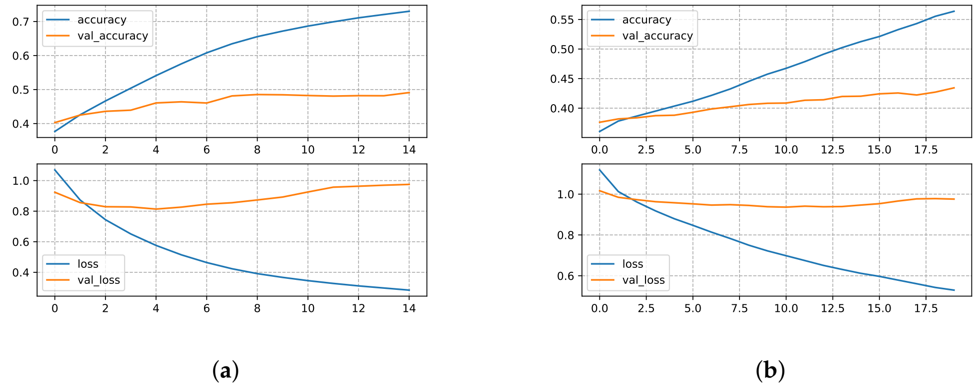


(**a**) (**b**)

Figure 5. Training and validation (`val_`) accuracies (upper plots), and training and validation (`val_`) losses (lower plots) for the Transformer trained with the CN+NCD KB: (**a**) N = 1, (**b**) N = 2.

In the case of the two-block model (Transformer 2) trained with the CN+NCD KB (Figure 5b), the training and validation losses dropped to around 0.475 bits and developed during a much larger number of training steps (we set 40 epochs max.) with an even smaller divergence (0.05 bits), showing thus much better generalization with respect to what we observed in the same model trained only with the OpenNCDKB. The same can be said for the accuracies reaching about ≈53%, that this time showed a tendency to keep improving with a small divergence (only about 1.6%), indicating stability between the train and validation predictions and the low variability of the model when it is exposed to unseen data (generalization). It is worth noting that the model improved throughout the 40 epochs we set as maximum, which suggests that the results could be improved by increasing the total number of epochs.

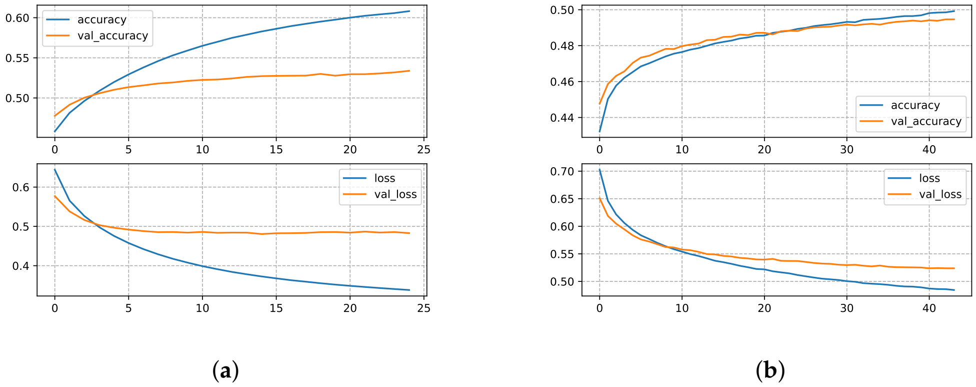
Like in cases analyzed before, the apparent lack of semantic information of our smaller KBs was generally highlighted by the results obtained from the SANLMs trained with them. Now training both Transformer 1 and Transformer 2 using the OIE-GP+NCD KB (36.54k training triples) exhibits similar behaviors than training them only with the OpenNCDKB (13k training triples). Although the OIE-GP KB is twice as large as OpenNCDKB, it is barely one-tenth the size of ConceptNet. The losses and accuracies shown in Figure 6a and 6b, for Transformer 1 and Transformer 2, respectively, indicate that there are no clear improvements in using the OIE-GP KB to enrich the OpenNCDKB. Therefore, 36.54k training triples are still an insufficient dataset size for our SANLMs to generalize in semantic reasoning tasks.



(**a**) (**b**)

Figure 6. Training and validation (`val_`) accuracies (upper plots), and training and validation (`val_`) losses (lower plots) for the Transformer trained with the OIE-GP+NCD KB: (**a**) N = 1, (**b**) N = 2.

Finally, the SANLMs trained using the ConceptNet+OIE-GP+NCD KB show similar results to those obtained with the ConceptNet+NCD KB (see Figure 7). Transformer 1 model (Figure 7a) required two less epochs compared to the same architecture trained on the ConceptNet+NCD KB. Training the Transformer 2 model (Figure 7b) with the ConceptNet+OIE-GP+NCD KB caused a decrease in performance, compared to training the model with the ConceptNet+NCD KB, i.e., 0.03 bits in the case of the loss and 0.01% in the case of accuracy. Despite this decrease, the model showed greater stability, as the train and validation curves were less divergent but showing increasing performance when the ConceptNet+OIE-GP+NCD KB was used.



(**a**) (**b**)

Figure 7. Training and validation (`val_`) accuracies (upper plots), and training and validation (`val_`) losses (lower plots) for the Transformer trained with the ConceptNet+OIE-GP+NCD KB: (**a**) N = 1, (**b**) N = 2.

7.2.2. Sensitivity Analysis for Attention-Based GRU Models

To understand the impact of hyperparameters on our attention-based GRU models' performance, we conducted a sensitivity analysis from 311 experiments total to evaluate their importance in influencing the loss function, which is minimized using the Bayesian optimization of the hyperparameters to optimize semantic reasoning for object phrase generation. In this analysis, we assessed which hyperparameters contribute the most to changes in the loss function using a random forest regression surrogate model, following the approach of Probst et al. [61]. Regarding hardware, we used four Nvidia RTX-4090 GPUs to train

311 models. We used Adafactor optimizer which adjusts the learning rate based on parameter scale. For this analysis we used our largest KB (the ConceptNet+OIE-GP+NCD KB).

The surrogate model predicts the loss function's value based on hyperparameter configurations, treating each hyperparameter as a feature. Feature importance scores are computed using Gini importance, which measures the total reduction in node impurity attributed to each hyperparameter across the random forest's trees. Additionally, we calculated Spearman's rank correlation coefficient to quantify the monotonic relationship between each hyperparameter's values and the loss function, providing insight into whether increases in a hyperparameter are associated with higher (worse) or lower (better) loss.

Table 4 presents the importance scores and Spearman correlation coefficients for the hyperparameters of our attention-based GRU models, sorted by importance in descending order. The table includes only variable hyperparameters, as batch size (64), number of epochs (5), and sequence length (30) were held constant to ensure stable training conditions. Importance scores range from 0 to 1, with higher values indicating greater influence on the loss function. Spearman correlations range from -1 to 1, where positive values indicate that higher hyperparameter values are associated with higher loss (and thus lower performance), and negative values suggest improved performance with higher values.

Table 4. Hyperparameter Importance and Correlation Analysis (sorted in descending order by importance score).

Hyperparameter	**Hyperparam. Values**	**Importance**	**Correlation**
Num. Layers	{1, 2, 4}	0.363	0.515
Dropout	[0.0–1.0] uniform	0.288	-0.095
Hidden Units	{512, 1024, 2048}	0.248	0.410
Embedding Dim.	{256, 512, 1024}	0.1	0.029
Batch Size	(64 constant)	—	—
Number of Epochs	(5 constant)	—	—
Seq. Len.	(30 constant)	—	—

The results in the table reveal that the number of layers has the highest importance score (0.363), indicating it is the most influential hyperparameter in determining the loss function's value. Its strong positive Spearman correlation (0.515) suggests that increasing the number of layers (from 1 to 4) is associated with higher loss, implying reduced model performance. This may indicate overfitting in deeper models, as additional layers increase complexity, potentially capturing noise in the OpenIE-derived medical knowledge base rather than generalizable patterns.

This trend is visually confirmed in the performance map (see Figure 8), where purple lines (low `val_loss`) cluster at lower `numLayers` values (e.g., 1.2–2.0), while yellow lines (high `val_loss`) are more frequent at higher values (e.g., 3.0–3.8). The dense packing of lines in these regions underscores the consistency of this pattern across numerous experiments. This implies that deeper models are prone to overfitting, capturing noise in the training data rather than generalizable patterns, which degrades performance on the validation set. For optimal generalization, shallower architectures (e.g., 1 or 2 layers) are likely preferable within this tested range.

Hidden units, with an importance score of 0.248 and a correlation of 0.410, also significantly affect the loss, with larger hidden unit sizes (e.g., 2048) leading to higher loss, possibly due to increased model capacity exacerbating overfitting on the noisy dataset. In Figure 8, larger `hiddenUnits` values may align with yellow lines, though specific clustering is not detailed.

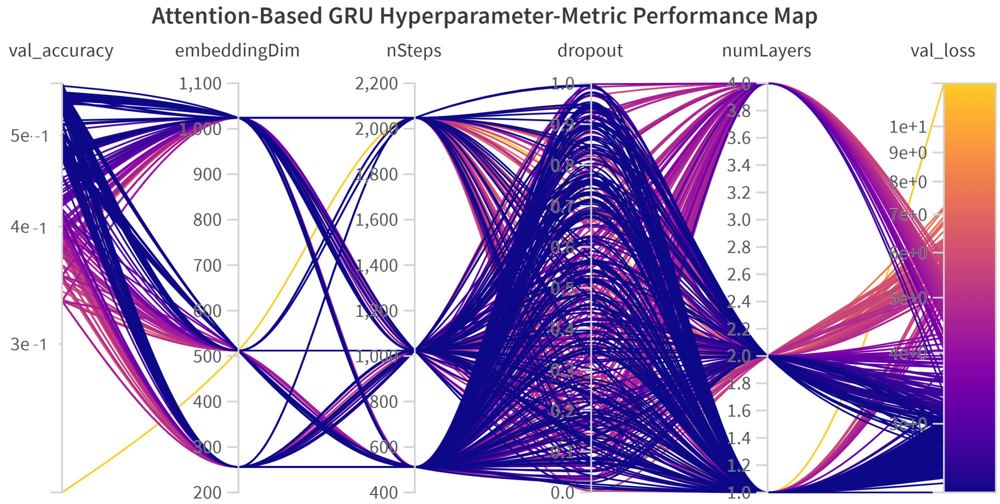


Figure 8. Global view of the experiments with attention-based GRU models, mapping hyperparameters to the loss function. Metrics are in logarithmic scale to ease visualization.

The dropout rate (`dropout`) is uniformly distributed over the range [0.0, 1.0], covering no regularization (0.0) to full regularization (1.0). Table 4 shows moderate importance score of 0.288, indicating a notable but secondary influence on `val_loss` compared to `numLayers`. Its Spearman correlation is weakly negative at -0.095, suggesting a slight tendency for higher dropout rates to reduce validation loss. The performance map shows a uniform distribution of dropout values across the full range, with a mix of purple and yellow lines throughout near to 0.5. This indicates no clear clustering of low `val_loss` (purple lines) or high `val_loss` (yellow lines) around this specific dropout rate, with weak and inconsistent effect on model performance. The regularization provided by dropout may mitigate overfitting to some extent as it tends to higher values, but its impact appears highly dependent on interactions with other hyperparameters, such as the number of layers.

The embedding dimension has the lowest importance (0.1) and a near-zero correlation (0.029), indicating minimal impact on the loss function. This suggests that the model's performance is relatively insensitive to changes in embedding size within the tested range (256 to 1024), possibly because the semantic reasoning task relies more on structural hyperparameters (e.g., layers, hidden units) than on embedding size.

The dropout rate (`dropout`) is uniformly distributed over the range [0.0, 1.0], covering no regularization (0.0) to full regularization (1.0). Table 4 shows moderate importance score of 0.288, indicating a notable but secondary influence on `val_loss` compared to `numLayers`. Its Spearman correlation is weakly negative at -0.095, suggesting a slight tendency for higher dropout rates to reduce validation loss. The performance map shows a uniform distribution of dropout values across the full range, with a mix of purple and yellow lines throughout near to 0.5. This indicates no clear clustering of low `val_loss` (purple lines) or high `val_loss` (yellow lines) around this specific dropout rate, with a weak and inconsistent effect on model performance. The regularization provided by dropout may mitigate overfitting to some extent as it tends towards higher values, but its impact appears highly dependent on interactions with other hyperparameters, such as the number of layers.

The embedding dimension has the lowest importance (0.1) and a near-zero correlation (0.029), indicating minimal impact on the loss function. This suggests that the model's performance is relatively insensitive to changes in embedding size within the tested range

(256 to 1024), possibly because the semantic reasoning task relies more on structural hyperparameters (e.g., layers, hidden units) than on embedding size.

These findings highlight the trade-offs in hyperparameter tuning for our task. While increasing model complexity (e.g., more layers or hidden units) can enhance expressive power, it risks overfitting, particularly in noisy, open-vocabulary knowledge bases derived from common sense knowledge and NCD literature. Conversely, regularization techniques like dropout offer a protective effect, though their impact, measured by means of correlation, is moderate. The low importance of embedding dimension suggests that computational resources may be better allocated to optimizing other hyperparameters. These insights guide model configuration, favoring simpler architectures with moderate regularization to balance performance and generalizability in medical knowledge reasoning.

In Figure 8, we have the global view of the experiments, complementing the aim of identifying critical hyperparameters and their effects on model generalization, particularly in the context of semantic reasoning tasks. We focus primarily on the number of layers (`numLayers`) and dropout (`dropout`), with additional insights into hidden units (`hiddenUnits`) and embedding dimension (`embeddingDim`).

7.3. Number of Layers

The number of layers (`numLayers`) takes discrete values of {1, 2, 4}, representing shallow to moderately deep architectures. According to Table 4, `numLayers` has the highest importance score of 0.363, indicating that it is the most influential hyperparameter in determining `val_loss`. Its Spearman correlation with `val_loss` is strongly positive at 0.515, suggesting that, as the number of layers increases from 1 to 4, the validation loss tends to rise significantly. This trend implies that deeper models are prone to overfitting, capturing noise in the training data rather than generalizable patterns, which degrades performance on the validation set. For optimal generalization, shallower architectures (e.g., 1 or 2 layers) are likely preferable within this tested range.

7.4. Dropout (Regularization)

The dropout rate (`dropout`) is uniformly distributed over the range [0.0, 1.0], covering no regularization (0.0) to full regularization (1.0). Table 4 assigns it a moderate importance score of 0.288, indicating a notable but secondary influence on `val_loss` compared to the number of layers. However, its Spearman correlation is weakly negative at −0.095, suggesting a slight tendency for higher dropout rates to reduce validation loss. This weak correlation implies that the dropout's effect is inconsistent across experiments, with no specific value within [0.0, 1.0] emerging as universally optimal. The regularization provided by dropout may mitigate overfitting to some extent, but its impact appears highly dependent on interactions with other hyperparameters, such as the number of layers.

7.5. Other Hyperparameters

While the number of layers and dropout are the primary focus, Table 4 provides insights into additional hyperparameters:

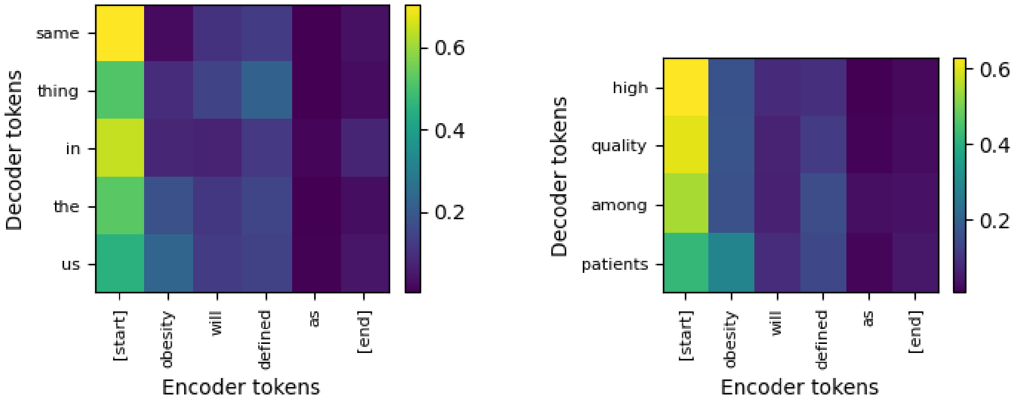
- Hidden units (`hiddenUnits`): With values {512, 1024, 2048}, it has an importance score of 0.248 and a positive correlation of 0.410 with `val_loss`. This indicates that larger hidden unit sizes increase validation loss, similar to `numLayers`, likely exacerbating overfitting in more complex models.
- *Embedding dimension (`embeddingDim`)*: Taking values {256, 512, 1024}, it has the lowest importance score of 0.1 and a near-zero correlation of 0.029. This suggests that `embeddingDim` has minimal influence on `val_loss`, making the model relatively insensitive to changes in embedding size within this range.

- *Constant hyperparameters*: Batch size (64), number of epochs (5), and sequence length (30) are held constant, with no importance or correlation data provided, indicating they were not varied in this analysis.

7.6. Attention Matrix

We see in Figure 9 that we have the attention matrices of best-performing models with high and middle regularization probabilities. Figure 9a shows the attention matrix of the best-performing model according to the validation loss (the global minimum among the 311 experiments).

The attention matrix of the model reveals how the model aligns encoder and decoder tokens to generate object phrases. The matrix shows attention weights between the encoder tokens (`[start], obesity, will, defined, as, [end]`) on the x axis and decoder tokens (`same, thing, in, the, us`) on the y axis, with a color gradient from purple (low attention, ≈0.0) to yellow (high attention, >0.6). The model's hyperparameters and metrics are `dropout = 0.765`, `embeddingDim = 512`, `nSteps = 512`, `numLayers = 1`, and `val_loss = 2.271`.



(**a**) Attention headmap of the best-performed-dropout model.

(**b**) Attention headmap of the mid-dropout model.

Figure 9. Attention matrix visualization of the best-performing models according to the validation loss performance metric.

The attention matrix of the model reveals how the model aligns encoder and decoder tokens to generate object phrases. The matrix shows attention weights between encoder tokens (`[start], obesity, will, defined, as, [end]`) on the x-axis and decoder tokens (`same, thing, in, the, us`) on the y-axis, with a color gradient from purple (low attention, ≈0.0) to yellow (high attention, >0.6). The model's hyperparameters and metrics are: `dropout = 0.765, embeddingDim = 512, nSteps = 512, numLayers = 1`, and `val_loss = 2.271`.

Mid attention (green, ≈0.5) from [start] to *us* or *the* suggests the model prioritizes the start token to initiate decoding, establishing context for the phrase. For example, a token that appears at [start] is critical because it sets the context for all subsequent processing. In addition, on the one hand, tokens marked explicitly for position (*in*)—or those acting as connectors (for example "as")—may serve a syntactic function and therefore receive less emphasis. On the other hand, higher attention weights to these kind of word could indicate that localization or purpose/role context is critical. High attention weights to `[start]` also mean that the model relies strongly on the opening context, especially if the token implies or carries thematic content or creative semantic cues. In this sense, we think that since the start of a sentence often defines the overall tone or gist, high attention here, and depending on how the model learned to do so, ensures that these salient cues are not diluted as the sequence unfolds.

Mid attention (green, ≈0.5) from [start] to *us* or *the* suggests that the model prioritizes the start token to initiate decoding, establishing context for the phrase. For example, a token that appears at [start] is critical because it sets the context for all subsequent processing. In addition, on the one hand, tokens marked explicitly for position (*in*)—or those acting as connectors (for example "as")—may serve a syntactic function and therefore receive less emphasis. On the other hand, higher attention weights to these kinds of word could indicate that localization or purpose/role context is critical. High attention weights to `[start]` also mean that the model strongly relies on the opening context, especially if the token implies or carries thematic content or creative semantic cues. In this sense, we think that since the start of a sentence often defines the overall tone or gist, high attention—depending on how the model learned to do so—here ensures that these salient cues are not diluted as the sequence unfolds.

Moderate-to-low attention (blue, ≈0.2–0.3) from *obesity* to *us* or *the* indicates the model links the key entity (*obesity*) to determinants followed by entity, suggesting in this case entity-centric (with respect to the U.S., as a country) and geographical reasoning for phrases like *in the us* in the context of defining obesity. Notice that the low (but >0.2) attention from *defined* to *thing*, which suggests that the model learns to link the verb with the generated (although generic) object. The model uses 1 layer, aligning with the sensitivity analysis's recommendation (importance 0.363, correlation 0.515) for shallow architectures. The high dropout = 0.765 (importance 0.288, correlation -0.095; higher dropout group > 0.625 also) regularizes the model, supporting the tendency observed in the importance-correlation analysis (higher regularization associated to lower validation loss) by preventing over-specialization on specific tokens.

We have now the attention matrix of the attention-based GRU model that performed the best within the mid-dropout group of models, i.e., dropout$\in [0.375, 0.625]$. It had the following configuration: `dropout = 0.613`, `embeddingDim = 1024`, `nEpochs = 5`, `nSteps (hidden units) = 512`, `numLayers = 1`, and `validation loss = 2.358`, visualized as a heatmap in Figure 9b. The matrix maps attention between encoder tokens (`[start], obesity, will, defined, as, [end]`) and decoder tokens (`high, quality, among, patients`). Compared to the previous attention matrix, this one shows mid to low attention from `obesity` to `patients`, and from *defined* to `among`. Due to this learned association, the decoder tokens (`patients, among, high, quality`) can jointly be related to the NCD domain, indicating the adaptation of the model to NCD phrases. This is in contrast with the generic previous tokens ("same thing in the us"), where the context barely resembles the location of the disease (according to training data). As in the previous case, this model learned to weigh the verb (*defined*) with the decoder sequence, but not with a noun this time. In this matrix, the same attention pattern is observed for `[start]` (to the first token in the decoder) and `[end]` (low attention to the ending token of the decoder) than that observed previously in the best-performing model.

Now, we have the attention matrix of the best performing model with a lower dropout group (<0.0375). See Figure 10. In this case, the dropout = 0.271, and configuration is `embeddingDim` = 256, `nSteps` = 1024, `numLayers` = 1, and `val_loss` = 2.409. We observe that, while in the mid-regularized model generic NCD-specific phrases (*patients, among, high, quality*) are generated, in the high-dropout (best-performing) and low-dropout (this) models, generic phrases are generated. Broader context is observed (*world*, *we*, *be*, *on*) for this low-dropout model (but best-performing), possibly indicating a shift to general semantic reasoning. Notice that, now, with a lower dropout, the lower embedding Dim = 256 (vs. 512 and 1024) may limit nuanced medical focus, while nSteps = 1024 broadens the learned scope (to the world). Similarly to the above mentioned patterns, *defined as* is mapped to the decoders phrase *world we*.

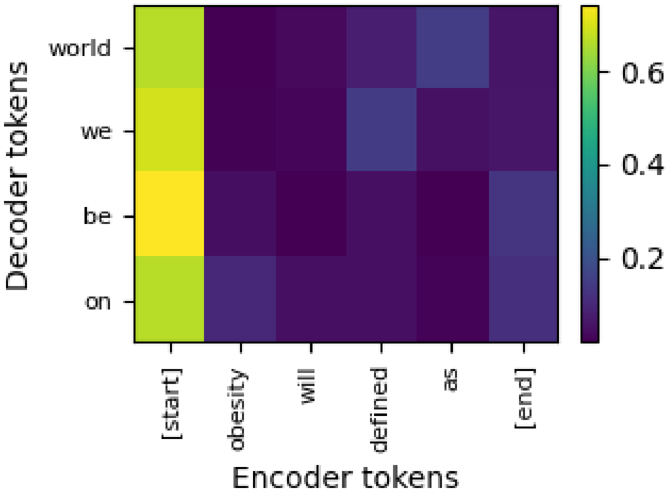
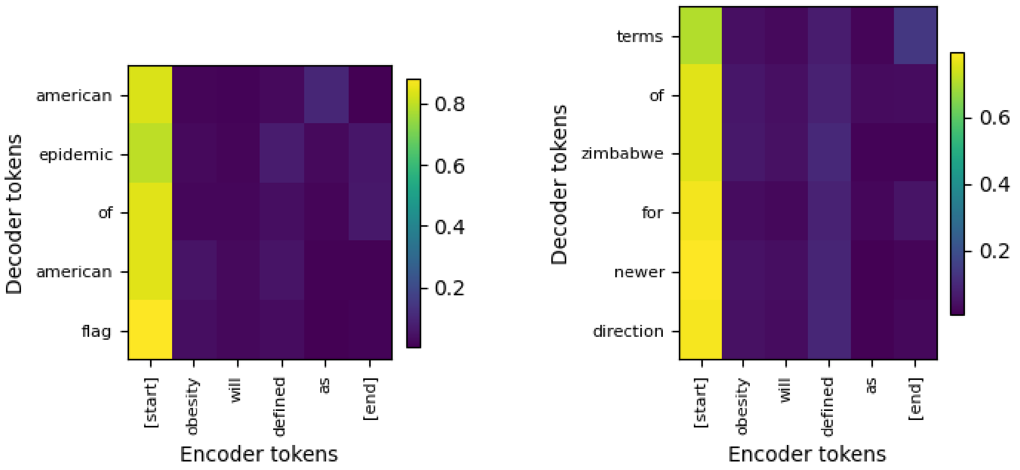


Figure 10. Attention matrix of the best-performing model in the low-dropout group.

7.7. Attention Matrix of the GRU-W Model

As we observed in the matrices analyzed thus far, only the best-performing mid-dropout model in Figure 9b generated a phrase that can be related (though generic) to medical domains. This result prompted us to revisit the Bahdanau-style configuration [42], which we label GRU-W (embedding dimension 256; hidden units 2048; one recurrent layer). GRU-W has twice as many hidden units (wide) but half the embedding dimensions of the mid-dropout model, reshaping the network's capacity rather than merely enlarging it. We therefore ask whether a setup proven effective for machine-translation tasks can also foster richer semantic reasoning. At first glance, our 311 experiments show that these configurations tend towards overfitting, due to the relatively large number of hidden units (the larger this number, the larger the loss, according to Table 4).

In Figure 11a, we have the GRU-W model with a high dropout = 0.834. The encoder tokens form the same phrase as before and the decoder tokens are *american*, *epidemic*, *of*, *american*, *flag*. The model shows very similar attention pattern with respect to most of the models analyzed at this moment from `[start]` and `[end]` tokens. The heat-map shows three salient off-diagonal peaks. Firstly, low attention from *obesity* to *american* suggests that the decoder re-uses the noun *obesity* to predict that the frame [X] is related to or characteristic of the United States, a plausible medical trope. Secondly, attention from *defined as* to *american epidemic*—low weights also—indicate lexical coupling through the verb, and repeating the adjective (*american*) can be a possible side-effect of noisy OpenIE tuples. Minimal weights on the positional token `[end]` reflect the heavy regularization; the model does not over-rely on boundary cues, consistent with the importance–correlation analysis in Table 4. Although semantically evocative ("American epidemic"), the model over-fits rare co-occurrences—possibly mirrored by its highest validation loss (2.422).



(**a**) (**b**)

Figure 11. Attention matrix visualization of the GRU-W configuration (**a**) with the best performance and (**b**) with the lower dropout and best performance according to the validation loss.

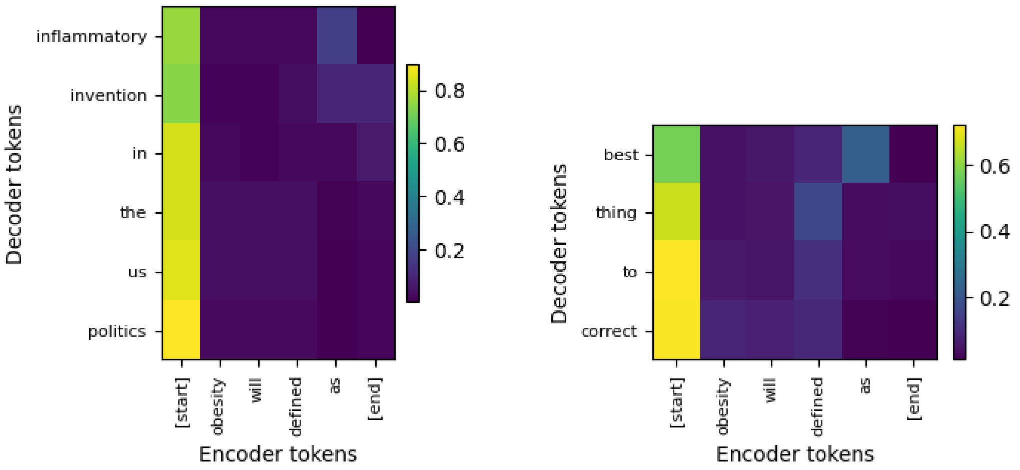
We have a more superficial interpretation without having access to more information (at this moment): the pattern shows focused attention on specific semantic relationships, with a thematic shift toward cultural or symbolic phrases, indicating content-driven generation over positional cues. More specifically, in this example, the model shows thematic reasoning by inferring that obesity is a societal issue ("American epidemic"), while also showing symbolic reasoning over NCD-specific focus, e.g., "obesity *is defined as a flag* of American epidemic". We think this can be due to the presence of Common Sense Knowledge in the training data. The `val_loss` = 2.422 is the highest among all to this point.

Now, we see that, in the low-dropout GRU-W model (dropout = 0.084, see Figure 11b), the graduated attention from `[start]`—starting at mid-high attention (terms), escalating to very-high attention (of, zimbabwe), and peaking at highest attention (for, newer, direction)—indicates a strong reliance on the start token to drive the entire decoding process, with increasing emphasis on later tokens that suggest purpose and progression.

Although there is information regarding the scientific community's concern regarding Zimbabwe's health crisis, we see a lack of disease entity-driven reasoning with obesity directly. Rather, the top-to-bottom sequence (terms, of, zimbabwe, for, newer, direction) suggests a phrase like "terms of Zimbabwe for a newer direction", possibly framing obesity management in a specific geographical and forward-looking context. The high attention from [start] to zimbabwe, for, newer, and direction supports this focus. The task's open-vocabulary nature allows for geographical references like Zimbabwe, and the model's focus on newer and direction aligns with a more general NCD literature's emphasis on innovation.

The lowest attention from *will*, *as*, and `[end]` (except very-low attention for *for* from `[end]`) indicates that these tokens have little influence on decoding, reinforcing the dominance of `[start]` and the lack of content-driven focus beyond defined's minor role. This pattern is consistent with prior matrices, where auxiliary tokens had minimal impact, though *defined*'s slight low attention to later tokens suggests a minor definitional role.

Given that, at this point we observed that GRU-W models showed improved entity-driven semantic reasoning, we decided to train separately two additional models with this configuration (number of hidden units is 2048 and embedding dimension of 256). This time, we used 10 epochs (instead of five like in the previous 311 experiments). In Figure 12, we see the attention matrices of the models with similar (to Figure 11) dropout (0.831, not equal due to random sampling) to that of the best performance (0.834), and with the lower dropout (0.012) and best performance according to the validation loss (3.583, and 2.927, respectively).



(**a**) (**b**)

Figure 12. Attention matrix visualization of independently trained models with the GRU-W configuration (**a**) with a similar dropout to that of the best performance, and (**b**) with the lower dropout and best performance according to the validation loss.

Figure 12a, re-run from Figure 11a, now with dropout = 0.831, shows generic NCD-specific entity focus. The attention from `[start]` varies from green (high) to *inflammatory* and *invention*, suggesting that the model uses `[start]` to initiate decoding with a focus on medical and innovative terms (*inflammatory*, *invention*) and increases with contextual connectors (*in*, *the*, *us*). Yellow (very high) for *politics*, and greener yellow (higher) for *in*, *the*, and *us*, indicates a strong positional influence on prepositional and national context tokens. The highest attention peak to *politics*, suggests that `[start]` drives the decoding toward a societal or political conclusion.

Read semantically from top to bottom, the decoder tokens suggest a phrase like "inflammatory invention in the US politics". In the NCD context, this could imply "inflammatory invention in the US politics related to obesity", focusing on a medical innovation within a national political framework. The lowest attention across most decoder tokens from *will* and *defined* indicates negligible impact from these tokens; low attention from *as* for *inflammatory* and very-low attention for *invention* suggest a minor influence in assigning roles to these medical terms. The lowest to very-low attention from *obesity* across all decoder tokens suggests that the model seems to neglect the central NCD entity.

Now, Figure 12b shows the re-run from Figure 11b, but now with dropout = 0.012 and 10 epochs. Moderately high attention is observed from `[start]` to *best*, suggesting that the model begins decoding with an evaluative term. Very-high attention from `[start]` to *thing* indicates a strong positional influence on a general noun. The highest goes from `[start]` to *to* and *correct*: This peak attention suggests that the start token drives the decoding towards a corrective action with strong structural emphasis. Again, the key NCD entity obesity has almost no influence on decoding, which is unexpected for the task. And again, low attention (but not the lowest) is assigned from *defined* to *thing*, and from *as* to *best*, indicating the certain importance of the definitional role of *best thing*, together with *to correct*. The NCD task often involves corrective strategies, and the model's focus on correct aligns with this goal, though the lack of obesity focus undermines its relevance.

Sensitivity Analysis for Attention-Based LSTM Models

To understand the impact of hyperparameters on our attention-based LSTM models' performance, we conducted a sensitivity analysis from a total of 168 experiments to evaluate their importance in influencing the loss function (`val_loss`). This analysis, complemented by Bayesian optimization, aimed to optimize semantic reasoning for object phrase generation. We assessed which hyperparameters contribute most to changes in the loss function. The findings were somewhat similar to those for the GRU models, with key differences influenced by a more constrained hyperparameter search space, so making our analysis more cost-effective. That is, we explored only one and two layers (compared to up to 4 for GRUs) and a discrete set of dropout values {0.01, 0.5, 0.9} (instead of continuous samples in [0, 1]). These modifications notably shifted the perceived importance of certain hyperparameters. Again, in this case, we used our largest KB (the ConceptNet+OIE-GP+NCD KB) and the same hardware as for the attention-based GRU models.

Table 5 presents the importance scores and Spearman correlation coefficients for the varied hyperparameters of our attention-based LSTM models, sorted by importance in descending order. Constant hyperparameters such as batch size (64), number of epochs (5), and sequence length (30) were not varied and are thus excluded from this part of the analysis. Importance scores (0 to 1) quantify the influence on the loss function, while Spearman correlations (-1 to 1) indicate the direction of this influence (positive correlation means that a higher hyperparameter value is associated with a higher loss/poorer performance).

Table 5. Hyperparameter importance and correlation analysis (sorted in descending order by importance score) for the attention-based LSTM models.

Hyperparameter	Hyperparam. Values	Importance	Correlation
Hidden Units	{512, 1024, 2048}	0.413	0.032
Num. layers	{1, 2}	0.259	0.311
Dropout	{0.01, 0.5, 0.9}	0.218	0.118
Embedding Dim.	{256, 512, 1024}	0.110	0.277
Batch Size	(64 constant)	—	—
Number of Epochs	(5 constant)	—	—
Seq. Len.	(30 constant)	—	—

The global view of these experiments is captured in Figure 13. In this plot, the lines connect hyperparameter values to performance metrics. The color of the lines is determined by `val_loss` (purple = low loss, indicating better performance; yellow = high loss, indicating poorer performance). The `val_accuracy` is explicitly shown on the leftmost axis, where higher values signify better performance. Ideally, purple lines (low loss) should correspond to high `val_accuracy`.

7.8. Hidden Units (`nSteps`)

The number of hidden units (labeled `nSteps` in Figure 13) has the highest importance score (0.413) concerning `val_loss`, making it the most influential hyperparameter in this regard. Its Spearman correlation with `val_loss` is very weakly positive (0.032).

Figure 13 illuminates the following: regarding `val_loss`, dark purple lines (lowest loss) are densely clustered when `nSteps` is around 512. As `nSteps` increases to 1024 and 2048, lines tend towards orange and yellow (higher loss). Regarding `val_accuracy`, the lines originating from `nSteps` = 512 predominantly lead to the higher end of the `val_accuracy` axis. Conversely, `nSteps` = 2048 shows many lines terminating at lower `val_accuracy` values. `nSteps` = 1024 is mixed but generally achieves lower accuracy than 512. This dual observation confirms that smaller hidden unit sizes (specifically 512) are optimal, leading to both lower loss and higher accuracy, likely by preventing overfitting.

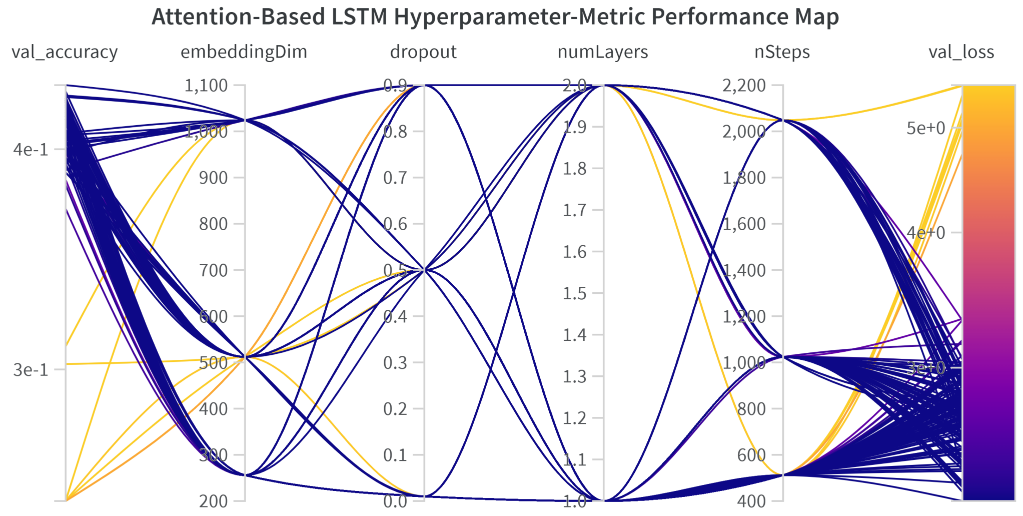


Figure 13. Global view of the experiments with attention-based LSTM models, mapping hyperparameters to the loss function. Metrics are in logarithmic scale to ease visualization. Notice that *e*-notation (e.g., `5e-1`) stands for scientific notation (e.g. 5×10^{-1}).

7.9. Number of Layers

The number of layers holds the second-highest importance score (0.259 for `val_loss`) with a moderate positive Spearman correlation of 0.311 with `val_loss`. The explored values were {1, 2}.

The impact is starkly visible in Figure 13. Regarding `val_loss`, `numLayers` = 2.0 is overwhelmingly associated with yellow and orange lines (high loss). `numLayers` = 1 shows many purple lines (low loss). Regarding `val_accuracy`, lines from `numLayers` = 1.0 predominantly lead to higher `val_accuracy` values. In contrast, lines from `numLayers` = 2 terminate at the lower end of the `val_accuracy` scale on several occasions.

This strongly indicates that a single-layer LSTM architecture is superior for this task, achieving both lower loss and higher accuracy. Two layers appear to induce severe overfitting, degrading both metrics.

7.10. Dropout

Dropout, with values of {0.01, 0.5, 0.9}, is the third in importance (0.218 for `val_loss`) and has a weakly positive Spearman correlation (0.118 with `val_loss`). Figure 13 provides critical insights. Regarding `val_loss`, the three regularization degrees explored are the source of a high concentration of dark purple lines (lowest loss). Also, yellow lines pass by the three values, which poses as the only conclusive evidence of the importance and correlation values mentioned for this hyperparameter.

This strongly indicates that a single-layer LSTM architecture is superior for this task, achieving both lower loss and higher accuracy. Two layers appear to induce severe overfitting, degrading both metrics.

7.11. Embedding Dimension

`embeddingDim` has the lowest importance score (0.110 for `val_loss`) but a moderate positive Spearman correlation (0.277 with `val_loss`), suggesting that larger dimensions tend towards higher loss. The explored values were {256, 512, 1024}. The addition of `val_accuracy` helps clarify its role.

Observing Figure 13, we see that, regarding `val_loss`, purple lines (low loss) are frequent for `embeddingDim` = 256 and 1024. `embeddingDim` = 512 has a greater share of orange/yellow lines (higher loss). The highest accuracy values are predominantly achieved with `embeddingDim` = 256 and 512. Lines from these values often reach the upper part of the `val_accuracy` axis. `embeddingDim` = 1024, while showing some spread, has lines leading to the lower or mid-range accuracy values and is associated with some of the poorest-performing models (low accuracy, high loss).

The `val_accuracy` axis reinforces the observation from `val_loss`: smaller to medium embedding dimensions (256 or 512) are preferable. They not only tend to result in lower loss in most cases but also in higher accuracy, suggesting that larger embedding dimensions might be adding unnecessary parameters without contributing to the better performance on this task.

These findings highlight that simpler LSTM architectures (1 layer, 512 hidden units), coupled with low-to-moderate regularization (0.01, 0.5 dropout) and reasonably sized embeddings (256–512), are key to maximizing the performance (both minimizing loss and maximizing accuracy) for the object phrase generation task. The `val_accuracy` metric provides a clearer picture of performance, especially for less dominant hyperparameters like `embeddingDim`, confirming the trends observed with `val_loss`.

7.12. Attention Matrix

To analyze the attention matrices of some representative LSTM models, we selected those that performed the best under certain conditions. First, in Figure 14, we have a heatmap of the attention matrix of the LSTM model that resulted in the lowest validation loss among the 168 experiments. This has a dropout = 0.9; embedding dim. = 256; number of hidden units = 512 and `val_loss` = 2.904. When generating the token *first*, the model primarily focused on the input word *defined*, and to a lesser extent on the beginning and end of the sequence (`[start]`, `[end]`). This suggests that *defined as* is a key context for initiating the generated object phrase with *first*. The token *life*, has very similar pattern, mainly differing in that less focus is observed at the end of the input sequence, but increased focus is observed for *defined as*. In general, the formed sentence is well formed, but senseless (*Obesity will [be] defined as first life of cars*).

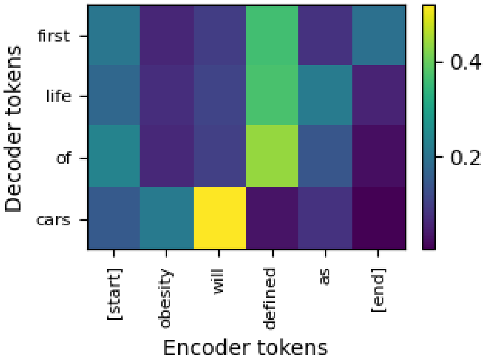
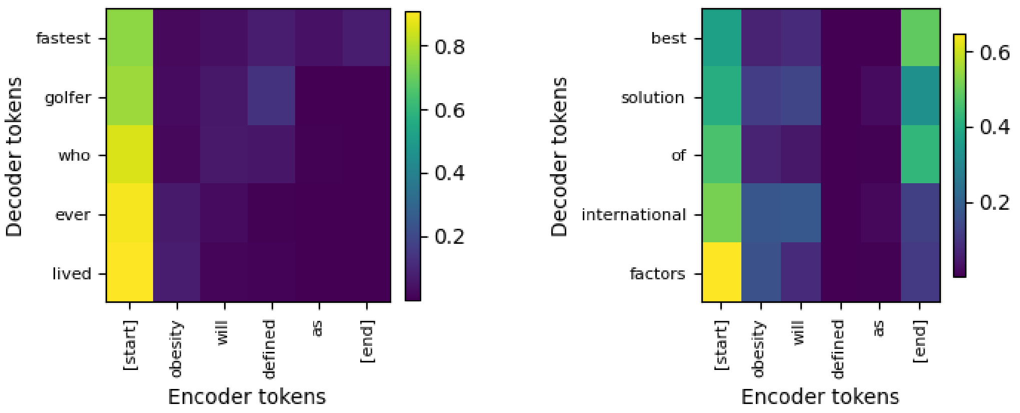


Figure 14. Attention matrix of the LSTM model that resulted in the lowest validation loss among the 168 experiments.

The next attention heatmap corresponds to the model that performed the best with the mid-dropout regularization (see Figure 15a): dropout = 0.5; embedding dim. = 256; number of hidden units = 2048 and `val_loss` = 2.674. It again generated a senseless object phrase, albeit showing normal attention patterns and an interesting small attention peak from the predicate to the adjective of the object phrase. The correspondence is structurally semantically sound, but with the wrong meanings selected.



(**a**) (**b**)

Figure 15. Attention matrix visualization of LSTM models (**a**) with mid-dropout (0.5) and (**b**) with the lower dropout (0.01). Best performances selected according to the validation loss.

Lastly, (in Figure 15b), we have the model that performed the best with the lowest dropout = 0.01; embedding dim. = 256; number of hidden units = 1024 (`val_loss` = 2.674). This model, although a generated senseless object phrase, it is nearer to the thematic context of its input. The object phrase makes reference to international context of obesity, which may be prompted by the attention of the decoder to the main thematic word in the encoder (*obesity*) and the future auxiliary (*will*). Notice, in addition, that this also represents an overgeneralization of the input context, such as it occurred with the GRU models.

7.12.1. MSPT on Best-Performed GRU and LSTM Models

In Table 6, we show the key hyperparameters resulting the best from the point of view of the validation loss (through the whole set of experiments) and with different degrees of dropout regularization: high, mid and low. In the case of the LSTM models, we set only three values: 0.9, 0.5, and 0.01. In the case of the GRU models, the values were grouped (high, mid, and low) from those available in the pool generated by Bayesian optimization using continuous range in [0, 1]. We labeled the models such that it is easy to identify their main hyperparameters in the subsequent analysis (see the second column of the table). For instance, `BestHdropSembMstep1L` stands for `Best` of all with `[H]`igh `[drop]`out, `[S]`mall `[emb]`edding dimension, `[M]`id amount of hidden (`[s]`teps) units and `[1 L]`ayer. We decided to include two-layered models here thinking about, in the event of that relying solely on the validation loss, we may miss interesting results from the point of view of our MSPT.

Table 6. Key hyperparameters of the GRU and LSTM models grouped by regularization degree. The best validation loss is marked in bold by the group and number of layers (1 and 2).

	Model Name	Dropout	EmbDim	nSteps	Num. Layers	Val Loss
GRU	`BestHdropMembMstep1L`	0.765	512	512	1	2.958
	`MdropSembHstep1L`	0.577	256	2048	1	5.502
	`LdropSembHstep1L`	0.084	256	2048	1	**2.869**
	`HdropSembHstep2L`	0.834	256	2048	2	2.960
	`MdropLembMstep2L`	0.613	1024	512	2	2.657
	`LdropSembMstep2L`	0.271	256	1024	2	**2.614**
LSTM	`BestHdropSembMstep1L`	0.90	256	512	1	**2.258**
	`MdropSembHstep1L`	0.50	256	2048	1	2.435
	`LdropSembMstep1L`	0.01	256	1024	1	2.302
	`HdropLembHstep2L`	0.9	1024	2048	2	2.702
	`MdropMembHstep2L`	0.50	512	2048	2	2.663
	`LdropMembHstep2L`	0.01	512	2048	2	**2.628**

In Table 7, we have the best-performig attention-based GRU models analyzed at this moment, but now seen from the lens of our MSPT. The best of them reached 2.959 bits of cross-entropy loss. It has 2048 hidden units, 256 embedding dimensions, one layer, and was regularized with 0.834 dropout probability (marked in bold). This model reached $\mu_{sts} = 0.582$ between its NCD test predictions and the corresponding gold standard, with a gap of 10.02% ($p < 3 \times 10^{-60}$).

An interesting result is that `NAtt` GRU models can even surpass the gap attained by attentional models in validation data, being comparable only to the 1-layered `LdropSembHstep1L` attentional model. The p-value of the best `NAtt` GRU model in test data is twenty magnitude orders higher than the best 1-layered attentional model. When think that this means that attentional models have more generalization ability to new domains, while `NAtt` models are better at the training domain.

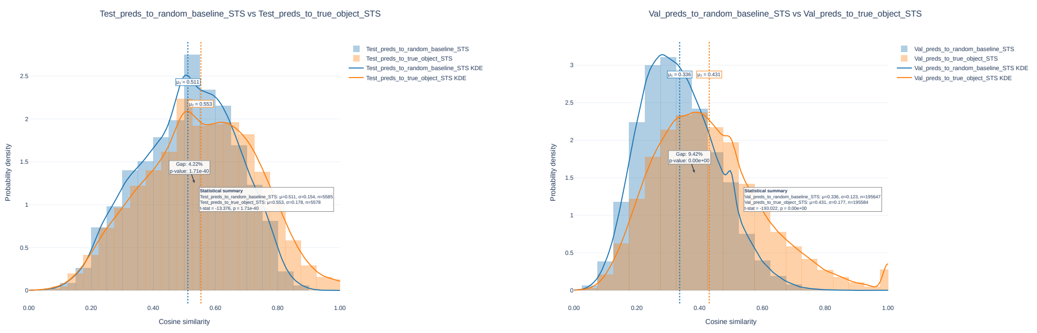
Table 7. Comparison of OpenNCDKB test and validation predictions of the attention-based GRU models. Results are ranked according to their *p*-value in test NCD data (OpenNCDKB). Best results are marked as bold.

Model Name	**Test NCD**				**Validation**			
	Pred	**Random**	**% Gap**	***p*-Value**	**Pred**	**Random**	**% Gap**	***p*-Value**
`HdropSembHstep2L`	**0.582**	0.529	**5.30%**	$\mathbf{3.57 \times 10^{-61}}$	0.440	0.339	10.10%	0.0
`LdropSembMstep2L`	0.572	0.522	5.00%	4.08×10^{-58}	0.436	0.339	9.70%	0.0
`MdropLembMstep2L`	0.581	0.535	4.60%	1.67×10^{-48}	0.438	0.342	9.60%	0.0
`LdropSembHstep1L`	**0.553**	0.511	**8.24%**	$\mathbf{1.71 \times 10^{-40}}$	**0.431**	0.336	**9.24%**	**0.0**
`BestHdropMembMstep1L`	0.576	0.536	4.00%	4.97×10^{-37}	0.435	0.341	9.40%	0.0
`MdropSembHstep1L`	0.442	0.438	0.85%	1.71×10^{-1}	0.376	0.347	8.41%	0.0
`NAttMdropLembSstep1L`	0.569	0.536	6.23%	2.69×10^{-27}	0.428	0.341	25.38%	0.0
`NAttHdropMembSstep1L`	0.561	0.541	3.70%	4.51×10^{-11}	0.415	0.341	21.81%	0.0

Figure 16 illustrates how our GRU-W model's cosine similarity scores differ when computed against truly matching object phrases versus random baseline phrases, on both the NCD test set (Figure 16a) and the validation set (Figure 16b). The blue histogram (and Kernel Density Estimation, KDE, curve) for the NCD test set shows similarities between model predictions and random baseline phrases: it peaks around ≈0.51 and is relatively wide, with very little (and vanishing) mass above ≈0.8. The orange curve for model-to-true object pairs is shifted slightly to the right (mean $\mu_{sts} \approx 0.53$) and has more density in the upper range (above 0.8). The vertical dashed lines mark these means and the annotated "Gap = 4.22%" (with $p \approx 1.71 \times 10^{-40}$) quantifies a small but highly significant advantage of matching over random.

For the validation set (Figure 16b), the random baseline distribution is centred lower ($\mu_{sts} \approx 0.33$) and remains tightly clustered with virtually no mass near 1.0. The true-object distribution sits higher ($\mu_{sts} \approx 0.43$), with a broader tail toward 1.0. The "Gap = 9.42%" (with p effectively 0) shows an even larger, statistically significant separation on the validation split.

In Table 8, we have the best-performing attention-based LSTM models analyzed at this moment, but now seen from the lens of our MSPT. The best of them reached 2.92 bits of cross-entropy loss. It has 2048 hidden units, 256 embedding dimensions, one layer and was regularized with 0.5 dropout probability (marked in bold). This model reached $\mu_{sts} = 0.584$ between its NCD test predictions and the corresponding gold standard, with a gap of 9.16% ($p < 3 \times 10^{-50}$). Regarding `NAtt` LSTM models, the tendency to learn the training domain is more pronounced, as they barely achieve significance in the test NCD domain, while an increased gap is observed in the validation data.



(**a**) Test predictions to random baselines STS (**b**) Validation predictions to true objects STS

Figure 16. Cosine–similarity distributions from our MPST metric for the best validation-test regularized GRU-W model (`LdropSembHstep1L`). (**a**) *NCD test set* model–random baseline STS distribution (blue), contrasted with model–gold phrase STS distribution (orange). (**b**) *Validation set* model–gold phrase STS distribution, contrasted with model–random baseline STS distribution.

Table 8. Comparison of OpenNCDKB test and validation predictions of the attention-based LSTM models. Results are ranked according to their *p*-value in test NCD data (OpenNCDKB). Best results are marked as bold.

Model Name	**Test NCD**				**Validation**			
	Pred	**Random**	**% Gap**	***p*-Value**	**Pred**	**Random**	**% Gap**	***p*-Value**
`MdropMembHstep2L`	**0.584**	0.535	**4.90%**	$\mathbf{3.67 \times 10^{-53}}$	0.442	0.342	10.00%	0.0
`HdropLembHstep2L`	0.576	0.533	4.30%	7.66×10^{-43}	0.440	0.345	9.50%	0.0
`BestHdropSembMstep1L`	0.575	0.541	3.40%	3.69×10^{-30}	0.430	0.344	8.60%	0.0
`LdropSembMstep1L`	0.572	0.537	3.50%	2.22×10^{-29}	0.427	0.341	8.60%	0.0
`LdropMembHstep2L`	0.560	0.531	2.90%	7.56×10^{-22}	0.423	0.342	8.10%	0.0
`MdropSembHstep1L`	0.481	0.480	0.10%	5.66×10^{-1}	0.371	0.354	1.70%	0.0
`NAttHdropSembMstep1L`	0.545	0.532	2.41%	1.92×10^{-5}	0.401	0.343	16.89%	0.0
`NAttMdropSembHstep1L`	0.531	0.523	1.50%	1.19×10^{-2}	0.397	0.343	15.88%	0.0

Argument for GRU-W Configuration Advantage

While the attention-based LSTM models analyzed at this point offer evidence of that their generated phrases are not purely random (best result was for `MdropMembHstep2L` reaching $\mu_{sts} = 0.584$ with 4.90% gap w.r.t 0.535 of the random baseline, $p < 3 \times 10^{-50}$), they generated senseless sentences, out of context with respect to the input subject and object phrases. Therefore, attention-based LSTMs serve as baseline references but are still a bit far away from the performance showed by attention-based GRU models in the different scenarios analyzed. Particularly the GRU-W configuration (EmbDim = 256, nSteps = 2048) offers a potential advantage over standard models due to its enhanced semantic flexibility, which we observed via the attention matrix analysis and our MPST (i.e. GRU-`HdropSembHstep2L` $\mu_{sts} = 0.582$ with 5.3% gap w.r.t 0.529 of the random baseline, $p < 3 \times 10^{-60}$). An additional and very important observation arises: the 1-layered GRU-`LdropSembHstep1L` model architecture resulted to have the best balanced result both, in validation and test KBs: test gap = 8.24% ($p \approx 1.71 \times 10^{-40}$) and validation gap = 9.24% (p = 0). See Table 7.

More specifically, 1-layered high dropout GRU-W generates evocative phrases like “American epidemic of American flag” and “inflammatory invention in the US politics”, indicating thematic and symbolic reasoning from Common Sense Knowledge and NCD literature, unlike the generic outputs of high (“same thing in the us”) and low (“world we be on”) dropout models. Its balanced regularization (e.g., Drop 0.834) and reshaped capacity (twice the hidden units, half the embedding dimension vs. mid-dropout) prevent over-specialization, supported by the sensitivity analysis (dropout importance 0.288, correlation −0.095). Extended training (10 epochs) improves decoding patterns, suggesting adaptability, though overfitting risks remain. The low dropout GRU-W (10 epochs) aligns with corrective NCD themes ("best thing to correct"), hinting at potential entity focus with refinement, unlike other models' narrower focus.

Table 9 compares the analyzed GRU models with key hyperparameters and attention performance insights. For instance, in the fisrt row we have `mid * [start]` → `us/the`; low * `obesity` ➝ `us/the`, where *mid* stands for mid attention weighting from the encoder `*`-associated token. That is, mid attention is assigned from the `[start]` token to the `us/the` tokens, and so on.

Table 9. Comparison of GRU models with key hyperparameters, main attention patterns, and output object phrase.

Model Group	**Drop**	**EmbDim**	**nSteps**	**Attn Patterns**	**Semantic**
Best High Dropout	0.765	512	512	mid * `[start]` → us/the; low * obesity → us/the	"same thing in the us"
Best Mid Dropout	0.613	1024	512	mid-low * obesity → patients; low * defined → among	"patients among high quality"
Best Low Dropout	0.271	256	1024	low * obesity; high * [start] → world	"world we be on"
GRU-W High Dropout	0.834	256	2048	low * obesity → american; low * defined/as → epidemic	"American epidemic of American flag"
GRU-W Low Dropout	0.084	256	2048	high * `[start]` → zimbabwe/for; low * obesity	"terms of Zimbabwe for newer direction"
GRU-W High (10 epochs)	0.831	256	2048	green * `[start]` → inflammatory/invention; highest * politics	"inflammatory invention in the US politics"
GRU-W Low (10 epochs)	0.012	256	2048	mid-high [start] → best; highest → to/correct	"best thing to correct"

7.12.2. Final Meaning-Based Selectional Preference Test (MSPT)

In this final MSPT, we first compared predictions of the attention-based GRU models using mixed KBs. First, an OpenIE-derived KB (OIE-GP+NCD, relatively small) and then this later including common sense knowledge (CN+OIE-GP+NCD, our largest KB). After that, we used these KBs along with an additional pair of small and large KBs (OpenNCDKB and CN+NCD) to train two transformer models from scratch (one- and two-block models). As introduced in Section 4.6, this evaluation takes into account three sources of information: the predicted (`pred`) and the ground truth object phrases, as well as the randomized ground truth object phrases (`rdn`, i.e., randomized gold standard or baseline). This latter serves to simulate a perturbation in selectional preferences, where the subject and predicate influence the generation of the object phrase.

This operation, as in previous sections, was performed on validation data and test data separately. The validation data refers to the validation subset of the training KB (e.g. CN+NCD). The test data is the test subset of our OpenNCDKB, which contains only NCD-related (noisy) OpenIE triples. This is for testing how the trained models behave in a purely (OpenIE-derived) biomedical domain considering the minimal vocabulary to which they were exposed during training.

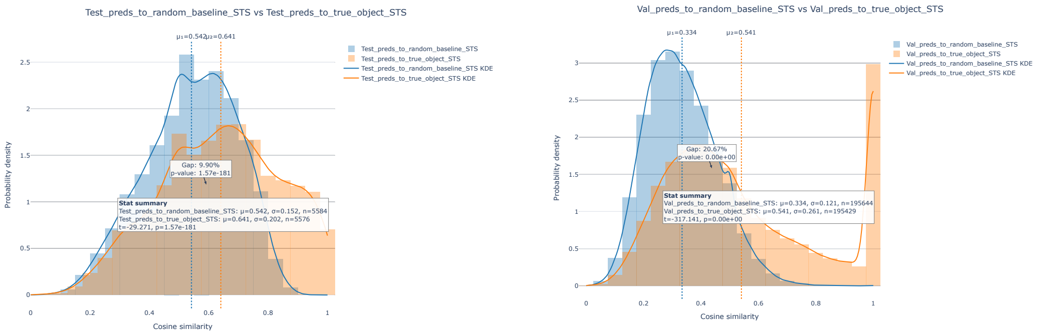
In Table 10 (penultimate row), we show our most significant recurrent model, this model architecture is supported by having the best result both in validation and test data (see `LdropSembHstep1L` in Table 7: dropout = 0.553; test gap = 8.24% ($p \approx$ 1.71e-40) and validation gap = 9.24% (p = 0.0). Now this model architecture (GRU-W) was evaluated using the CN+OIE-GP+NCD KB for which we had the the best global result. With this, the null hypothesis can be safely rejected with probability $1 - 1.57 \times 10^{-181}$ (gap 0.641-0.542→12.1%), which notably didn't require dropout to attain its performance. Its mean similarity with respect to the ground truth in the NCD test KB was 0.641 ($p \sim 0.0$), while in the validation KB it was 0.541, with gap 0.541-0.334 → 20.7%.

Table 10. p-values obtained using source and target test KBs to evaluate the MSPT confidence of the attention-based GRU SANLMs predictions (bold p-value indicates the most confident model and training KB).

Model and KB		Validation		Test NCD	
		μ_{sts} **(pred/rdn)**	p	μ_{sts} **(pred/rdn)**	p
GRU-W	OIE-GP+NCD	0.516/0.358	0.0	0.428/0.429	0.554
	CN+OIE-GP+NCD	0.541/0.334	0.0	0.641/0.542	$\mathbf{1.57 \times 10^{-181}}$
GRU-S	OIE-GP+NCD	0.474/0.361	0.0	0.502/0.499	0.267
	CN+OIE-GP+NCD	0.408/0.344	1.0	0.442/0.442	0.844

Based on these indicators, we believe that the GRU-W SANLM generates object phrase meanings and selectional preferences quite similar to those in the test and validation KBs. However, our analysis of the learning curves for this model (Figure 3a) shows that the model generates different words than the ground truth. Nonetheless, when considering both analyses together and the nature of the sentence embeddings used to represent the model predictions, we think that semantic selectional preferences with respect to the ground object phrases are strongly present in the meanings generated by the GRU-W model. This, we think, can be considered a manifestation of *generalization through semantic relatedness and selectional preference* in semantic reasoning.

In Figure 17, we see the cosine similarity distributions and their KDEs for the MPST of the GRU-W model trained on the CN+OIE-GP+NCD KB. Figure 17a shows that the STS distribution of the *NCD test set* pairs made between model predictions and random baseline phrases (blue bars and KDE) form a bell–shaped density centred at 0.542 ($\sigma = 0.152$, $n = 5584$) and drop sharply beyond 0.8. When the same predictions are compared with their gold object phrases (orange), the density flattens and shifts rightward, accumulating probability near 1.0 and yielding a higher mean of 0.641 ($\sigma = 0.202$). The 9.9% relative gap is highly significant ($t = -29.271$, $p < 1 \times 10^{-180}$).



(**a**) Test predictions to random baselines STS

(**b**) Validation predictions to true objects STS

Figure 17. Cosine–similarity distributions from our MPST metric for the GRU-W model.(**a**) *NCD test set* Model–random baseline STS distribution (blue), contrasted with model–gold phrase STS distribution. (**b**) *Validation set* Model–gold phrase STS distribution, contrasted with model–random baseline STS distribution.

In Figure 17b we see the *validation set* MSPT. The pattern persists in general: model–gold similarities remain right-shifted ($\mu_{sts} \approx 0.541$) and pile up close to 1.0, whereas model–random baseline similarities contract further and centre at 0.334 (gap $\approx$ 20.7%; $t = -317.141$, $p = 0.0$). Together, both figures therefore confirm that the model aligns far better with the true object phrases than with random baselines across both data splits.

Now, in Table 11, we show a comparison of the Transformer models' predictions using MSPTs between the predicted and ground truth object phrases (`pred`), and between the predicted and randomized ground truth object phrases (`rdn`). Our most significant model resulted when using the CN+NCD KB for training the Transformer-1, so the null hypothesis can be safely rejected. The mean similarity with respect to the ground truth in the NCD test KB was 0.551 ($p{\sim}0.0$) with gap of 4% w.r.t the corresponding random baseline of 0.511 ($p{\sim}1.33 \times 10^{-25}$), while in the validation KB it was 0.414. Such model required one Transformer block (encoder/decoder) and 20 training epochs. This performance was followed by the remaining Transformer-1 models trained with the CN+OIE-GP+NCD ($p{\sim}1.32 \times 10^{-22}$ in the test KB), OpenNCDKB, and OIE-GP+NCD KBs.

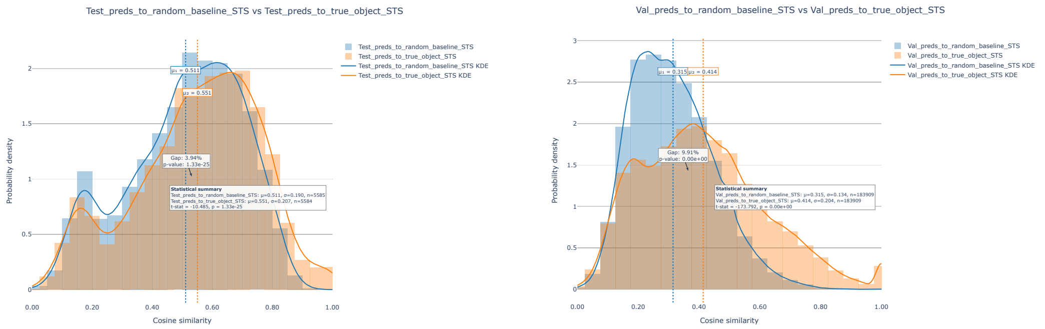
Table 11. p-values obtained using validation and test KBs to Evaluate the confidence of Transformer SANLMs transfer predictions (bold p-values indicate the most confident models and training KBs).

Model and KB		Validation		Test NCD	
Transformer-1	CN+OIE-GP+NCD	0.414/0.309	0.0	0.506/0.468	1.32×10^{-22}
	CN+NCD	0.414/0.315	0.0	0.551/0.511	$\mathbf{1.33 \times 10^{-25}}$
	OIE-GP+NCD	0.459/0.404	5.16×10^{-179}	0.584/0.569	1.95×10^{-8}
	OpenNCDKB	0.610/0.595	3.31×10^{-7}	0.613/0.596	2.83×10^{-11}
Transformer-2	CN+OIE-GP+NCD	0.342/0.298	0.0	0.424/0.418	0.0708
	CN+NCD	0.367/0.310	0.0	0.500/0.492	**0.0468**
	OIE-GP+NCD	0.284/0.284	1.0	0.308/0.308	1.0
	OpenNCDKB	0.604/0.604	1.0	0.607/0.607	1.0

In the case of the two-block Transformer models, most of their predictions on the NCD-related validation data can be confidently regarded as random (the null hypothesis holds). Even those predictions of the model trained with our largest KB (CN+OIE-GP+NCD). From these Transformer-2 models, only the one trained with CN+NCD proved to be significantly reliable, but with relatively low mean similarity w.r.t. the ground truth on the validation data ($\mu_{sts}0.367$, $p = 0.0$). Based on the corresponding performance analysis (Section 7.2.1), which showed that the two-block models generalized well but underperformed on the test data, we believe that their low-confidence MSPT measurements in the validation predictions (and therefore for the test set) may be due to task-level overfitting, potentially attributed to model size, as smaller models like GRU-W performed much better. Despite this, the predictions of the one-block models trained with CN+OIE-GP+NCD and CN+NCD were the most confident in distinguishing from the random baseline in test data.

The only two-block model that resulted significant barely reached $p < 0.05$, which was trained with the CN+NCD KB. These models trained with the smaller KBs (OIE-GP+NCD and OpenNCDKB) evidenced their poor performance (which initially appeared to be generalization), as their predictions can be reliably discarded as random (regarding meanings and selectional preferences) in both the test and validation data ($p{\sim}1.0$). Note that the MSPT results for these models reveal only modest gaps between each prediction and its shuffled counterpart (`pred-rdn`). The largest gap in validation data, $0.414 - 0.309 = 0.105$ (10.5%), is obtained with the second–ranked transformer—the one-block models trained on CN+OIE-GP+NCD. In contrast, the best dropout-regularized GRU attains a 5.3% gap ($p < 3 \times 10^{-60}$). It is worth mentioning that the Transformer-1 model trained with the O IE-GP+NCD KB is interesting because it reached relatively good evaluation results (cross-entropy loss and MSPT), both in test and validation data, despite the small size of the training data, which is infeasible to the GRU models.

Figure 18 illustrates how the Transformer-1 model's cosine-similarity scores differ when computed against ground truth object phrases versus random baselines, on both the NCD test set (a) and the validation set (b). In the case of the *NCD test set*, the blue histogram (and KDE) for model–random baseline STS peaks at $\mu_{sts} \approx 0.511$ ($\sigma \approx 0.19$, $n = 5585$) and falls off sharply beyond 0.8 with no mass near to 1.0. In contrast, the orange histogram for model–true-object pairs shifts rightward to $\mu_{sts} \approx 0.551$ ($\sigma \approx$0.207), accumulating more density beyond STS > 0.8. The vertical dashed lines mark these means, and the annotated "Gap = 3.94%" ($t = -10.485$, $p \approx 1.33 \times 10^{-25}$) quantifies a modest but highly significant advantage of correct over random matches. For the validation set, random-baseline similarities (blue) are tightly clustered at a lower mean of $\mu_{sts} \approx 0.315$ ($\sigma \approx 0.134$, $n = 183{,}909$), with virtually no mass near 1.0. True-object similarities (orange) rise to $\mu_{sts} \approx 0.414$ ($\sigma \approx$0.204), displaying a pronounced tail toward (and peaking again) 1.0. Here the "Gap = 9.91 %" ($t \approx -173$, $p \approx 0$) indicates an even larger, statistically significant separation. Together, both panels confirm that Transformer-1 predictions align substantially better with true object phrases than with random baselines across both splits.



(**a**) Test predictions to random baselines STS (**b**) Validation predictions to true objects STS

Figure 18. Cosine–similarity distributions from our MPST metric for the Transformer-1 model. (**a**) *NCD test set* Model–random baseline STS distribution (blue), contrasted with model–gold phrase STS distribution. (**b**) *Validation set* Model–gold phrase STS distribution, contrasted with model–random baseline STS distribution.

Together, these MSPT density plots reveal not only the numerical gaps between random baselines and true-object similarities but also how each model behaves when it is trained primarily on commonsense knowledge yet evaluated on the domain-specific NCD corpus. In particular, GRU-W produces a similarity distribution that is more tightly clustered and shifted toward 1.0 for true-object phrases—and more narrowly concentrated at lower values for random baselines—than the Transformer-1 model.

Overall, our MSPTs indicated that only the largest KBs provided the necessary knowledge for the SANLMs to consistently and confidently distinguish themselves from the random baseline in generating and selecting meanings. We interpret this as high stability in the semantic reasoning decisions made by the models. The best-performing models generated meanings that were consistently related to and selected in concordance with the ground truth. While this does not completely rule out the possibility of overfitting, the results of our MSPT-based evaluation method suggest that cross-entropy is effective for training the models but not necessarily for evaluating them from the perspective of semantic reasoning.

7.12.3. Manual Inspection of Common Sense Knowledge Predictions (Qualitative Error Analysis)

In the following manual inspections, Selectional Preferences—the tendency of a predicate to favor certain types of arguments—have been indirectly evaluated through the identification of frame shifts, semantic role misalignments, domain generalization failures. Specifically, we have examined whether the generated predictions by the models maintain the expected argument structure and intended communicative function of the predicate from the gold standard or if they introduce implausible or unrelated arguments.

In Table 12, we show predictions of the most significant baseline model GRU-W for five subject-predicate randomly sampled from the CN+OIE-GP validation KB. Consider the statement "music is a form of communication" as the Ground Truth. The prediction in question is “music is a form of sound.” When analyzing the predicate “is a,” we recognize it activates a hypernymy frame. This requires the object of the sentence to specify a category that adequately describes the subject. In this case, the subject is “music.”

Table 12. Five samples taken from the CN+OIE-GP test KB and used for inference using the most significant attention-based GRU model.

1	music is a [start] form of communication [end]	
	CN+OIE-GP+NCD (GRU-W)	[start] form of sound [end]
2	ride a bicycle motivated by goal [start] enjoy riding a bicycle [end]	
	CN+OIE-GP+NCD (GRU-W)	[start] the car to be safe [end]
3	car receives action [start] propel by gasoline or electricity [end]	
	CN+OIE-GP+NCD (GRU-W)	[start] train to jump [end]
4	an Indian restaurant used for [start] selling Indian meals [end]	
	CN+OIE-GP+NCD (GRU-W)	[start] curry style [end]
5	mediators capable of [start] settle a disagreement [end]	
	CN+OIE-GP+NCD (GRU-W)	[start] beat word [end]

In terms of roles, “music” (the Theme) should align with a hypernym that fits its abstract and communicative nature. For the Ground Truth, "communication" is a fitting hypernym, as it reflects the socially meaningful and expressive function of music. However, the prediction "music is a form of sound" violates this alignment. "Sound" is a broader, more physical-category hypernym that does not capture music’s communicative purpose effectively.

7.12.4. Manual Inspection of Common Sense Knowledge Predictions (Qualitative Error Analysis)

In the following manual inspections, Selectional Preferences—the tendency of a predicate to favor certain types of arguments—have been indirectly evaluated through the identification of frame shifts, semantic role misalignments, and domain generalization failures. Specifically, we examined whether the generated predictions by the models maintain the expected argument structure and intended communicative function of the predicate from the gold standard or if they introduce implausible or unrelated arguments.

In Table 12, we show predictions of the most significant baseline model GRU-W for five subject–predicate samples randomly sampled from the CN+OIE-GP validation KB. Consider the statement “music is a form of communication” as the ground truth. The prediction in question is “music is a form of sound”. When analyzing the predicate “is a”, we recognize it activates a hypernym frame. This requires the object of the sentence to specify a category that adequately describes the subject. In this case, the subject is “music”.

In terms of roles, “music” (the theme) should align with a hypernym that fits its abstract and communicative nature. For the ground truth, “communication” is a fitting hypernym, as it reflects the socially meaningful and expressive function of music. However, the prediction “music is a form of sound” violates this alignment. “Sound” is a broader, more physical category than hypernym that does not capture music’s communicative purpose effectively.

The Ground Truth “car receives action [start] propel by gasoline or electricity [end]” specifies a relevant action for a car. The prediction “train to jump” is nonsensical and unrelated to the function of a car. The frame activated by “receives action” requires an action performed on the subject that is relevant to its function. The Ground Truth satisfies the constraints by specifying “propel by gasoline or electricity” as a relevant action for a car. However, the prediction violates these constraints by introducing “train to jump,” which is irrelevant and nonsensical for a car. This results in a semantically incoherent prediction that fails to meet selectional preferences.

The gold standard ("music is a form of communication") satisfies the necessary constraints, indicating that "communication" is an appropriate category for music. Conversely, the prediction ("music is a form of sound") overgeneralizes due to hypernymic bias, failing to meet the selectional preferences for contextual specificity. Therefore, the prediction exhibits partial plausibility: it is semantically related but overgeneralizes, resulting in a failure to capture the specific communicative essence of music.

In the second example, the ground truth "ride a bicycle motivated by goal "enjoy riding a bicycle" specifies a hedonic motive for the action. The prediction "the car to be safe" introduces a functional goal that is not aligned with the activity of riding a bicycle. The frame activated by "motivated by goal" requires a purpose or motive that fits the activity's enjoyment or purpose.

The gold standard satisfies the constraints by using "enjoy" as a hedonic motive, fitting the activity of riding a bicycle. However, the prediction violates these constraints by introducing "to be safe" as a functional goal and incorrectly assigning "car" as the object of "ride". This results in a mismatch from the semantic roles point of view and reflects lexical substitution bias. Therefore, the prediction is nonsensical, failing to capture the hedonic motive and introducing an irrelevant object.

The ground truth "car receives action [start] propel by gasoline or electricity [end]" specifies a relevant action for a car. The prediction "train to jump" is nonsensical and unrelated to the function of a car. The frame activated by "receives action" requires an action performed on the subject that is relevant to its function. The ground truth satisfies the constraints by specifying "propel by gasoline or electricity" as a relevant action for a car. However, the prediction violates these constraints by introducing "train to jump", which is irrelevant and nonsensical for a car. This results in a semantically incoherent prediction that fails to meet selectional preferences.

In example 4, the ground truth "an Indian restaurant used for selling Indian meals" specifies the primary function of an Indian restaurant. The prediction "curry style" is vague and does not specify the restaurant's primary function. The frame activated by "used for" requires a function or purpose related to the subject's primary purpose. The gold standard satisfies the constraints by explicitly describing the primary function of the restaurant—serving and selling food. However, the prediction is too vague and does not clearly convey the function of an Indian restaurant, merely suggesting an aesthetic or general characteristic without specifying its role. Therefore, the prediction exhibits partial plausibility: it is semantically related but too vague, failing to capture the specific function of the restaurant.

In example 5, the ground truth "mediators capable of settle a disagreement" specifies a relevant ability for mediators. The prediction "beat word" is nonsensical and unrelated to mediation. The frame activated by "capable of" requires an ability or capacity related to the subject's role. The ground truth satisfies the constraints by specifying "settle a disagreement" as a relevant ability for mediators. However, the prediction violates these constraints by introducing "beat word", which is nonsensical and unrelated to mediation. This results in a semantically incoherent prediction that fails to meet selectional preferences.

In Table 13, we present the predictions of the most significant Transformer models for five subject-predicate concatenations randomly sampled from the ConceptNet KB. In general, all generated outputs maintain structurally valid sentence formations from a syntactic and grammatical perspective. However, Transformers tend to produce more fluent yet formulaic noun phrases—e.g., *"a form of expression"* versus the GRU-generated *"form of sound"*—possibly reflecting differences in attention mechanisms.

Table 13. Five samples taken from the ConceptNet test dataset and used for inference using the specified Transformer models.

1	music is a [start] form of communication [end]	
	CN+OIE-GP+NCD (Transformer 1)	[start] a form of communication [end]
	CN+NCD (Transformer 1)	[start] a form of expression [end]
	CN+NCD (Transformer 2)	[start] a form of expression [end]
2	ride a bicycle motivated by goal [start] enjoy riding a bicycle [end]	
	CN+OIE-GP+NCD (Transformer 1)	[start] you are tired [end]
	CN+NCD (Transformer 1)	[start] you need to get somewhere [end]
	CN+NCD (Transformer 2)	[start] you like to play [end]
3	car receives action [start] propel by gasoline or electricity [end]	
	CN+OIE-GP+NCD (Transformer 1)	[start] drive by car [end]
	CN+NCD (Transformer 1)	[start] drive by car [end]
	CN+NCD (Transformer 2)	[start] find in car [end]
4	an indian restaurant used for [start] selling indian meals [end]	
	CN+OIE-GP+NCD (Transformer 1)	[start] eat a meal [end]
	CN+NCD (Transformer 1)	[start] eating [end]
	CN+NCD (Transformer 2)	[start] eating [end]
5	mediators capable of [start] settle a disagreement [end]	
	CN+OIE-GP+NCD (Transformer 1)	[start] eat cat food [end]
	CN+NCD (Transformer 1)	[start] be [UNK] [end]
	CN+NCD (Transformer 2)	[start] be [UNK] [end]

For the case of *"music is a"*, Transformer 1 (trained with CN+OIE-GP+NCD) correctly predicted the gold-standard object phrase *"a form of communication"*, satisfying selectional preferences. However, when trained on CN+NCD (N = 1 and N = 2), the same model instead generated *"a form of expression"*. While the latter may sound pragmatically appropriate in the context of the arts, replacing *"communication"* with *"expression"* introduces semantic misalignments that alter the original intent of the statement.

From a Frame Semantics perspective, the gold standard activates a *communication frame*, where music operates as a structured system for meaning transmission. The models' output, however, activates an *expression frame*, which does not necessarily entail interaction or message interpretation. This shows that the model overgeneralizes the function of *music* beyond its original communicative intent. Models trained on CN+NCD substitute "expression", broadening the category to subjective output (e.g., personal creativity). While "expression" is semantically related, it violates the SP for interactive meaning transmission, resulting in partial plausibility.

In the next example, "ride a bicycle motivated by goal [enjoy riding a bicycle]", the models replace the emotional concept (positive emotion) of *"enjoying"* an activity (riding a bicycle) with a simpler idea conveying the purpose which a bicycle is conceived for (its functional definition). By testing Frame Semantics here, instead of framing cycling as an enjoyable activity, it suggests fatigue as a primary factor—potentially implying that cycling is a consequence of being tired rather than an activity performed for enjoyment. This activates an unintended exhaustion frame, rather than an intrinsic motivation or recreational activity frame. The case of Transformer 1 (CN+NCD): *"you need to get somewhere"* and "like to play" (Transformer 2 CN+NCD) are similar by shifting to transport and recreation frames, respectively. These violate SP constraints: "motivated by goal" requires intentionality tied to the activity itself, not external factors (fatigue) or unrelated actions (play).

Although in this later example the Transformer 2 model (CN+NCD) partially retained the idea of enjoyment (*"to play"*), it actually misrepresents the activity as a form of play rather than an intentional recreational or fitness activity. The phrase *"like to play"* broadens the interpretation beyond cycling, making the meaning more ambiguous. While cycling can be playful, the gold-standard phrase refers specifically to the enjoyment of cycling itself, not general play. More specifically, the term *play* may be age-dependent, commonly associated with children rather than adults cycling for recreation or fitness. This adds an unintended layer of interpretation (i.e., pragmatic misalignment), potentially distorting real-world applicability such as AI systems predicting user behaviors or preferences.

A similar pattern is observed in the following sample (3), *"car receives action [propel by gasoline or electricity]"*. From a syntactic and grammatical perspective, all generated outputs are structurally valid and fluent. However, they semantically diverge from the gold standard, leading to significant shifts in meaning and function. The gold standard emphasizes that the car's movement is powered by a specific energy source—a crucial detail for understanding its operational mechanism. The predicate "receives action" activates, for example, a mechanical propulsion frame, where the car (Patient) is acted upon by an energy source (Instrument).

The gold standard satisfies selectional preferences with "gasoline/electricity" as valid energy sources in the corresponding frame. All predictions violate this later: in the case of "drive by car" (CN+OIE-GP+NCD and CN+NCD, Transformer 1), although the model invokes a driving frame—typically associated with the human action of operating a vehicle, it incorrectly assigns the car as Agent (implying self-propulsion). This makes the predictions semantically incompatibles with the subject–predicate input. Notice also that "find in car" (CN+NCD, Transformer 2) introduces an unrelated location frame.

In the fourth example, the ground truth states "an Indian restaurant used for [selling Indian meals]". The predicate "used for" activates a commerce frame, requiring the object to specify the restaurant's (Theme) transactional functioning. The gold standard assigns "selling meals", (Function) satisfying SP by emphasizing economic exchange, which is what the restaurant is used for. Predictions ("eat a meal", "eating") shift to a consumption frame, changing the restaurant's role to a consumer action, which can occur in various contexts and does not capture the specificity of the food service industry. While semantically related (meals are eaten), this violates selectional preferences by ignoring the commercial function (selling), resulting in the partial plausibility of the formed sentences.

Entry number five shows distinct failures in generating a contextually appropriate object phrase. In this case, the model was prompted by the phrase *"mediators capable of"*—which, in theory, should trigger an object related to the mediators' (Agent) role in the conflict resolution frame in which the ground truth agrees with "settle disagreement" (Action). Instead, the models generated *"eat cat food"*. Although the Action role is maintained and agrees with the predicate, this output clearly deviates from the intended function of mediators frame. This is a clear violation of selectional preferences, as the predicted verb does not fit the expected argument type.

The issue appears to arise because the model relies on statistical co-occurrence rather than proper semantic role labeling. The phrase *"capable of"* should have prompted an output indicating the mediators' ability to *"settle a disagreement"* or perform a related conflict resolution task. However, the model instead generated *'eat cat food'*, likely due to common associations in the training data where *'eat'* is frequently linked with routine consumption actions, even if the resulting pairing (like with cat food) is unusual.

Moreover, while the generated verb *"eat"* might seem to be treated as the main verb, it is important to note that it was generated solely as the output corresponding to the input *"mediators capable of"*. The error, therefore, lies in the model's failure to link the modifier *"capable of"* with the correct domain-specific action. Instead of reinforcing a conflict resolution frame, the output triggers an irrelevant eating frame.

Note that, in general, the random samples analyzed so far showed that the two-block models did not learn semantic agreement between the elements of the semantic structures, even when they managed to distinguish themselves from the random baseline according to our MSPT. This suggests that, to be more confident about the knowledge acquired by the model, it must pass the MSPT with high significance on both the test and validation data.

7.12.5. Manual Inspection of Biomedical Predictions (Qualitative Error Analysis)

In Table 14, we have the predictions of the GRU-W model on test data from the OpenNCDKB. See that, in the first example, there is a plausible substitution made by the model by using *"with various roots"* instead of *"to clinical variables"*. Although the two phrases are not semantically related on their own, they are semantically related when looking at the subject–predicate phrase provided to the model (*"these subtypes are closely correlate"*), even when no specific subject noun is present, i.e., *"these subtypes"*. We observe that the model uses the contexts associated to *"are closely correlate"* in the NCD-related data to perform its inference, the most probable word sequence. The model overuses frequent noun phrases (*"various roots"*), failing to keep the correlation frame by shifting to non-clinical, ambiguous referent frame unrelated to medical discourse.In addition, the themes linked by the model, *"subtypes"*, and *"root causes"* result in semantically nonsensical sentence because selectional preferences are violated in that the model ignored that predicate *"correlated"* requires co-varying quantitative entities.

Table 14. Five samples randomly chosen from the inferences made over the test OpenNCDKB using the most significant attention-based GRU model.

1	these subtypes are closely correlated [start] to clinical variables [end]	
	CN+OIE-GP+NCD (GRU-W)	[start] with various roots [end]
2	Immune - based interventions are opening up [start] entirely novel therapeutic avenues [end]	
	CN+OIE-GP+NCD (GRU-W)	[start] independent potentially treatment venture [end]
3	glycemic targets should reflect [start] presence of comorbid conditions [end]	
	CN+OIE-GP+NCD (GRU-W)	[start] for the association between dietary factors including chronic hyperglycemia in the evening [end]
4	these factors may provide [start] an opportunity to reduce deaths [end]	
	CN+OIE-GP+NCD (GRU-W)	[start] a combination of factors such as age gender cardiovascular risk factor [end]
5	Ethnic minorities to have [start] worse control of their diabetes [end]	
	CN+OIE-GP+NCD (GRU-W)	[start] all their care in the primary care setting [end]

In the second example, the generated object phrase has a grammatical error (*potentially* → *potential*). In addition, the prediction *"independent potentially treatment venture"* is nonsensical due to the frame being shifted by replacing the medical innovation frame (induced by the predicate *"are opening up"*) with business venture. Also, *"venture"* could not fulfill the Result role in a medical context, which leads to role misassignment. Therefore, selectional preferences were violated because the model failed to meet semantic constraints for *"opening up"* by generating a non-medical, but commercial term.

The third example involves more specific concepts, i.e., *"glycemic targets"* and *"comorbid conditions"* in the subject and object phrases, respectively. The generated sentence states that *"Glycemic targets should reflect for the association between dietary factors, including chronic hyperglycemia in the evening"*, which has no major syntactic or grammatical errors. The prediction presents frame shift as it replaces medical adjustment with the epidemiological association frame induced by the predicate (*"reflect"*). Also, *"dietary factors"* and *"hyperglycemia"* cannot fulfill the Factor role in a clinical guideline context, so there is a role misassignment. Therefore, semantic preferences are violated because the model fails to meet constraints for the predicate, which requires its object to be a clinically actionable factor (e.g., comorbid conditions like hypertension or kidney disease) that directly influences treatment decisions.

The fourth prediction *"a combination of factors such as age gender cardiovascular risk factor"* is semantically implausible in the context of the subject and predicate. This is because the model replaces opportunity for intervention frame with factor enumeration. Also, "Combination of factors" cannot fulfill the "Outcome" role in the ground truth, thereby showing role misassignment. Thus, the resulting sentence is nonsensical as the model fails to meet constraints for *"provide"* being an outcome (bust be an actionable opportunity, e.g., *"reduce deaths"*) by generating non-actionable and descriptive terms (age/gender/cardiovascular risk), which are input variables instead.

The fifth prediction *"all their care in the primary care setting"* is also semantically implausible. This is because the predicate *"have"* activates the Health Disparity frame, emphasizing inequitable health outcomes (e.g., poor diabetes control) linked to marginalized groups. However, the model shifts to a Healthcare Logistics frame (*"primary care setting"*), focusing on care delivery rather than health outcomes. Also, the theme "Ethnic minorities" appropriately serves as the group experiencing inequity. However, the model misassigns the attribute role to a logistical descriptor, stripping the sentence of its equity focus, i.e., *"Worse diabetes control"* fulfills the role of a measurable, inequitable outcome. Therefore, the model failed to meet constraints for *"have"* because *"Care in the primary care setting"* describes a process, not an outcome.

Now, we analyze the performed predictions on five randomly selected triples in the noncommunicable diseases domain (OpenNCDKB) using Transformer models. These are listed them in Table 15. In the case of the inputs (1) and (4), the models predict the most likely and grammatically well-formed phrases, e.g., *"cardiovascular disease"* given *"subtypes closely correlated"*. This is not factual at all as there is no a concrete subject in the input (*"these subtypes"* and *"these factors"*). The two-block model generated many more free constructions for these inputs.

Table 15. Five samples randomly chosen from the inferences made over the OpenNCDKB test set using the most significant Transformer models.

1	these subtypes are closely correlated [start] to clinical variables [end]	
	CN+OIE-GP+NCD (Transformer 1)	[start] with a [UNK] [end]
	CN+NCD (Transformer 1)	[start] with cardiovascular disease [end]
	CN+NCD (Transformer 2)	[start]with the development of the management of the disease[end]
2	Immune - based interventions are opening up [start] entirely novel therapeutic avenues [end]	
	CN+OIE-GP+NCD (Transformer 1)	[start] to treat the disease [end]
	CN+NCD (Transformer 1)	[start] [UNK] to cure individuals [end]
	CN+NCD (Transformer 2)	[start] to the development of the disease [end]
3	glycemic targets should reflect [start] presence of comorbid conditions [end]	
	CN+OIE-GP+NCD (Transformer 1)	[start] the risk of gastric cancer [end]
	CN+NCD (Transformer 1)	[start] the risk of microvascular complications such as retinopathy in patients with type 2 diabetes mellitus [end]
	CN+NCD (Transformer 2)	[start] the [UNK] of the [UNK] of the [UNK] [end]
4	these factors may provide [start] an opportunity to reduce deaths [end]	
	CN+OIE-GP+NCD (Transformer 1)	[start] the risk of lung cancer [end]
	CN+NCD (Transformer 1)	[start] more effective in treatment options in lung cancer [end]
	CN+NCD (Transformer 2)	[start] the [UNK] of the [UNK] [end]
5	Ethnic minorities to have [start] worse control of their diabetes [end]	
	CN+OIE-GP+NCD (Transformer 1)	[start] the risk of lung cancer [end]
	CN+NCD (Transformer 1)	[start] more likely to be high among patients [end]
	CN+NCD (Transformer 2)	[start] the treatment of gastric cancer [end]

In the second input, *"Immune-based interventions are opening up"* (2), we observe behavior similar to that of the previously analyzed GRU-W model. The models generate generic phrasing with simpler terms—such as *"to treat the disease"* and *"to cure individuals"*—instead of the gold standard's *"entirely novel therapeutic avenues"*. This tendency to avoid more creative or domain-specific compounds was consistently observed across the test data.

Specifically, Transformer 1 (CN+OIE-GP+NCD) fails to capture the notion of novelty inherent in the gold standard. Rather than emphasizing new therapeutic approaches, it suggests a standard treatment objective (*"to treat the disease"*), thereby losing critical domain-specific details. In contrast, Transformer 1 (CN+NCD) exhibits a frame shift and conceptual error by generating *"[UNK] to cure individuals"*. Here, the model should have generated an output reflecting the novelty and expansion of therapeutic possibilities. Instead, it produces a generic curative action, losing the sense of innovation. Transformer 2 (CN+NCD) further misaligns the meaning with *"to the development of the disease"*, which completely inverts the intended message (from the gold standard) by implying disease progression rather than pioneering therapeutic strategies. This semantic drift results in a pragmatic misalignment that fails to capture the transformative impact of immune-based interventions.

In the case of the input *"glycemic targets should reflect"*, the gold standard establishes a broad medical principle by linking glycemic targets to the *"presence of comorbid conditions"*, emphasizing individualized treatment adjustments based on coexisting health factors. The Transformer models violated selectional preferences, as the expected argument type for the predicate in the input (*"reflect"*) was ignored because it typically takes arguments related to measurable or guiding factors for treatment (a clinically actionable factor). The Transformer 1 model trained with the CN+OIE-GP+NCD KB instead predicted *"the risk of gastric cancer"*. This represents a frame shift from the original *"Medical Treatment Planning"* frame to a *"Risk Assessment"* frame, altering the intended focus. Additionally, the model fails to preserve the correct semantic roles: while the theme (*"glycemic targets"*) and model predicate (*"should reflect"*) are implied in the ground truth, the model's prediction instead introduces an incorrect attribute (*"the risk of gastric cancer"*), i.e., a non-actionable factor that does not directly inform glycemic target adjustments. While diabetes can influence cancer risk [62], this prediction disrupts the intended message by failing to capture the adaptive nature of glycemic target recommendations.

The Transformer 1 model trained with the CN+NCD KB predicted *"the risk of microvascular complications such as retinopathy in patients with type 2 diabetes mellitus"*, which this time resulted differently than it did with previous predictions. While this output remains within the diabetes domain and is demonstrated to be valid medical knowledge [63], it overspecifies details rather than generalizing appropriately. The model shifts the focus from a medical adjustment frame to a more specific diabetes-specific complication frame (microvascular complications in diabetes). This results in a semantic misalignment, as the model retains the attribute role but introduces unnecessary roles (example role: *"such as retinopathy"*, Experiencer role: *"patients with type 2 diabetes mellitus"*) while still omitting the theme and predicate roles. The model's output does not generalize excessively but rather overcommits to a specific medical scenario, reducing applicability across different patient contexts. Therefore, selectional preferences are in conflict because the predicate *"reflect"* requires arguments that are quantifiable and relevant to treatment decisions (e.g., comorbidities like hypertension or kidney disease), not descriptive, or redundant details (e.g., specific complications). The Transformer 2 model was unable to generate a coherent output, producing a sequence of unknown tokens. This represents a complete failure in meaning generation, preventing any valid frame from being formed and resulting in a prediction that is semantically unusable.

The analysis of the fifth input *"Ethnic minorities to have worse control of their diabetes"* reveals that the models struggle to maintain the intended focus on health disparities related to diabetes management. Instead, they introduce either entirely different diseases or vague likelihood statements that fail to capture the original meaning. The first prediction, *"the risk of lung cancer"*, shifts the focus from disparities in diabetes control to a *disease risk assessment* frame, replacing the original attribute (*worse control of diabetes*) with a health outcome (*risk of lung cancer*). The verb *"have"* suggests possession or experience, but the generated phrase *"treatment of gastric cancer"* is not an appropriate argument under its selectional preferences. This substitution suggests that the model is relying on statistical co-occurrences in the training data rather than preserving structured reasoning about health disparities. The shift may also indicate bias amplification, as lung cancer is frequently discussed in medical literature, potentially reinforcing an unintended link between minorities and cancer.

The second prediction, *"more likely to be high among patients"*, lacks specificity and fails to form a coherent argument structure. The attribute role is left vague, and the theme (*ethnic minorities*) is missing entirely, making the statement ambiguous and disconnected from the intended meaning. This suggests a failure to establish a meaningful frame alignment, preventing the model from producing a clear assertion about disparities.

The third prediction, *"the treatment of gastric cancer"*, introduces yet another frame shift—this time, from *health disparities* to *medical treatment*, entirely replacing the original attribute role (*"worse control of . . . "*) with an unrelated concept (*"treatment of gastric cancer"*). Selectional preferences are also violated because *"have"* requires a health outcome, but *"treatment of gastric cancer"* is a clinical action, not an outcome. This not only distorts the intended message but also introduces another instance of bias reinforcement, as the model again links minorities to a severe illness that was never present in the input.

Across all three cases, the models fail to preserve key semantic roles, either omitting the theme (*ethnic minorities*) or replacing the intended attribute (*worse diabetes control*) with related medical concepts. The frequent frame shifts and role assignments indicate that the models are not reasoning about disparities in a structured way but are instead defaulting to generalized medical associations from the training data. This behavior ultimately leads to semantic drift and bias reinforcement, distorting the intended meaning and producing outputs that do not align with the original purpose of the input.

8. Position Statement on Noise Propagation Due to Pseudo-Labels

Our work establishes a foundational framework that serves as an ideal testbed for future research on noise propagation in knowledge-based reasoning systems. While the current implementation incorporates basic noise mitigation strategies through filtering techniques, we position this framework as particularly valuable for the systematic study of error propagation mechanisms—a critical area in both knowledge engineering and language modeling research.

8.1. Our Framework as a Testbed for Noise Propagation

The integration of OpenIE-derived knowledge bases with neural language models creates a unique experimental setting for studying how noise propagates through semantic reasoning systems. Unlike traditional knowledge bases with manually curated assertions, our approach deliberately operates on automatically extracted triples that contain varying degrees of noise—making it well-suited for investigating pseudo-label noise dynamics. As noted by Romero and Razniewski [4], knowledge representation learning systems face fundamental challenges when integrating information from heterogeneous and potentially contradictory sources. Our framework allows researchers to trace how initial extraction errors, such as non-factual triples (e.g., "diabetes, is, hot"), affect downstream reasoning capabilities, providing a controlled environment for studying this phenomenon.

To acknowledge potential errors at the semantic level, we perform the manual inspection of triples using frame semantics, roles, and selectional preferences, identifying invalid relations (e.g., "diabetes, is treated with, chocolate") that deviate from expected medical contexts. While not explicitly tracing errors to their exact source (e.g., specific text or extraction tool), this inspection serves as a preliminary form of tracking wrong knowledge by pinpointing semantic inconsistencies. Similarly to how Chen et al. [64] demonstrated the ability to trace reasoning paths in neural networks for fact verification, our framework potentially enables tracing reasoning paths to determine how noisy inputs influence object phrase generation. This capability makes it particularly valuable for developing techniques to identify and mitigate error propagation in symbolic–neural integration systems [65,66], which Bosselut et al. [22] identified as a significant challenge in commonsense knowledge acquisition. Future work could enhance this by improving our attention matrix interpretations (Section 7.2.2) and our Meaning-Based Selectional Preference Test (MSPT) to systematically quantify and trace error sources, such as unreliable extractions or misaligned semantic roles [67].

8.2. Connection to Hallucinations in Language Models

The issue of noise propagation in our framework directly parallels the hallucination problem in large language models—where models generate plausible-sounding but factually incorrect information [68]. Our framework has the potential to provide a controlled environment for studying hallucination mechanisms for several reasons:

1. Traceable knowledge sources: Unlike black-box language modeling and NLP multitask scenarios, our framework allows for the precise tracking of which knowledge triples contribute to a particular prediction, enabling researchers to distinguish between hallucinations, stemming from noisy data (e.g., "diabetes, is treated with, chocolate") versus model limitations (semantic reasoning model capacity driven by architecture hyperparameters). Manual inspection using frame semantics further supports this by identifying triples that violate selectional preferences.
2. Quantifiable distortion: As demonstrated by Ji et al. [69], hallucinations often represent distortions of training data rather than complete fabrications. Our framework, augmented by MSPT, enables measuring the semantic drift between input knowledge and generated outputs.
3. Controllable knowledge injection: Similarly to Lewis et al. [70], our approach allows for the controlled manipulation of the underlying knowledge, facilitating experiments on how different types and degrees of noise contribute to hallucination-like behaviors.

The observed behaviors in our system—where models sometimes generate plausible but factually unsupported object phrases—mirror the hallucination phenomena in LLMs, but with greater transparency regarding the source of errors. This characteristic makes our framework particularly suitable for developing and evaluating mitigation strategies that could eventually transfer to larger language models.

Future work can leverage this framework to explore targeted research questions around error propagation and hallucination mitigation, including (1) quantifying the relationship between input noise levels and hallucination frequency; (2) developing architectural modifications that enhance robustness to noisy pseudo-labels; and (3) exploring how different knowledge representation schemes affect error propagation dynamics. An improved MSPT could directly address error tracing by incorporating metrics for triple reliability or source attribution, building on frame-based validation techniques [67]. As noted by Farquhar [71] and Filippova [72], effective testbeds for hallucination research require both controlled variability and realistic complexity—criteria our framework satisfies through its integration of real-world OpenIE extractions with neural generation capabilities.

9. Conclusions

We trained different encoder–decoder SANLMs to perform semantic reasoning. On the one hand, the performance results (accuracy and loss) provided that the one-block model (Transformer 1) showed the best test metrics compared to the attention-based GRU-W model (low dimensionality with high order recurrence), i.e., 53% vs. 0.18%, respectively. However, train and test metrics showed divergence, which was interpreted as overfitting at first glance. On the other hand, the two-block model (Transformer 2) showed much more stable and much less divergent (although decreasing) performance metrics, which was interpreted as a generalization at first glance.

From the point of view of object phrase meaning generation, hypothesis testing based on MSPT revealed with high confidence, on the one hand, that the (non-regularized) GRU-W model reached the highest semantic relatedness generalization compared to the Transformer 1 model in test NCD data, i.e., $\mu_{sts} = 0.54$ ($p = 4.36 \times 10^{-181}$) vs. $\mu_{sts} = 0.41$ ($p = 1.35 \times 10^{-25}$), respectively. Notably, GRU-W without regularization reached the largest and highest significant gap with respect to the random baseline (>20%). Overall, these models successfully distinguished themselves from simulated perturbation of selectional preferences (the random baseline) without even learning directly from predicate constraints. This also confirmed that the models' reasoning generalization might be manifested in semantic displacements and logical implications (as well as the abstraction and simplification of meanings) rather than in the ground truth distribution of output tokens.

On the other hand, both the two-block Transformer model and the high-dimensionality, low-order recursion (fewer hidden units) GRU-S model ended up memorizing the training data, thus generating object phrases with almost random meanings in test experiments. Only one two-block Transformer model barely reached the significance required: Val $\mu_{sts} = 0.367$; $p < 0.05$.

Regarding manual inspection of prediction samples, we found that GRU-W models tested on common sense knowledge predictions struggle with overgeneralization (e.g., *"sound"* replacing *"communication"*) and syntactic errors. These models misalign semantic roles and neutralize affective intent (e.g., replacing *"enjoy"* with functional goals). Their outputs prioritize statistical associations over contextual precision. Transformers produce fluent but formulaic phrases (e.g., *"expression"* vs. *"communication"*), narrowing meanings while avoiding nonsensical errors. They partially preserve domain coherence (e.g., *"drive by car"*) but favor generic language over nuanced intent, reflecting training data biases. Formal analysis of selectional preferences showed that, while the predictions resulted in related concepts, in most cases, the constraints for frames and roles proposed by the predicates were not met.

Similarly, in biomedical contexts, GRU-W generates vague outputs (e.g., *"primary care"* vs. diabetes control disparities) and misassigned roles (e.g., symptoms as comorbidities). They overgeneralize domain knowledge (e.g., hyperglycemia $\neq$ comorbidity) and neutralize critical discourse, eroding clinical specificity. Transformers amplify biases (e.g., minorities linked to cancer risk) and overextend frames (e.g., *"risk of"* replacing actionable guidance). While avoiding GRU-W's vagueness, they overspecify details (e.g., retinopathy) or produce incoherent outputs (unknown token [UNK]). Larger models invert meanings (e.g., *"novel therapies"* → *"disease progression"*), prioritizing technical fluency over semantic frame accuracy.

Overall, both architectures prioritize grammatical safety and statistical frequency over contextual grounding, necessitating domain-specific training and semantic constraints to address overgeneralization, bias, and improve selectional preferences. The GRU-S and Transformer 2 models completely failed to learn all these skills, even when this latter model barely achieved $p < 0.05$ in our MSPT.

MSPT served as a quantitative guide to select the models that we analyzed manually. Notice that, even when the p-values resulted highly significant, the manual analysis revealed multiple fails in semantic reasoning, which is explained by the fact that the mean similarities between the predicted and ground truth object phrases were small in general (0.414 for the most significant Transformer model, and 0.541 for the most significant GRU model).

Computational resources is a major constraint that we think can be improved by means of evaluating models at the semantic level. Therefore, contributions to more complex but cost-effective models can be evaluated using more specialized metrics like MSPT where semantics is the main performance target.

In addition, it is also necessary to reason about predicate phrases and subject phrases, with the goal of editing and generating more diverse KBs. Although single-input–single-output sequence (encoder-decoder) models adapted to semantic reasoning tasks offer promising results, we believe that the need to propose new semantically motivated architectures in which each constituent phrase in the tuples of a KB is explicitly considered is becoming increasingly evident.

Finally, it is also needed to conclude that our preprocessing approach is limited, as it involves stop word filtering and annotation-based selection only for OpenNCDKB, balances data quality with resource efficiency, a critical consideration for medical domains like NCD literature. By avoiding extensive manual curation, we enable the rapid construction of knowledge bases from unstructured text, simulating real-world scenarios where resources are constrained. This noisiness, while introducing some challenges, allowed us to evaluate the robustness of basic attentional models in handling imperfect data, contributing to sustainable AI research.

Author Contributions: Conceptualization, I.A.-F.; Software, J.A.S.-R., A.T.-V., R.C.-B., I.G.-P. and I.A.-F.; Validation, I.A.-F.; Formal analysis, Y.I.B.-M. and I.A.-F.; Investigation, I.A.-F.; Resources, Y.I.B.-M.; Data curation, Y.I.B.-M., F.J.M., E.G.-R. and I.A.-F.; Supervision, I.A.-F.; Funding acquisition, I.A.-F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by SECIHTI grant number CF-2023-I-2854. The APC was funded by Instituto Nacional de Enfermedades Respiratorias Ismael Cosío Villegas.

Data Availability Statement: Our datasets and code are publicly archived at https://github.com/iarroyof/open-ncd-kbc (accessed on 2 June 2025).

Acknowledgments: We thank to Red Neuronal Lab members for the given support.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Harper, L.; Campbell, J.; Cannon, E.K.; Jung, S.; Poelchau, M.; Walls, R.; Andorf, C.; Arnaud, E.; Berardini, T.Z.; Birkett, C.; et al. AgBioData consortium recommendations for sustainable genomics and genetics databases for agriculture. *Database* **2018**, *2018*, bay088. [CrossRef] [PubMed]
2. Jehangir, B.; Radhakrishnan, S.; Agarwal, R. A survey on Named Entity Recognition—datasets, tools, and methodologies. *Nat. Lang. Process. J.* **2023**, *3*, 100017. [CrossRef]

3. Xu, Y.; Kim, M.Y.; Quinn, K.; Goebel, R.; Barbosa, D. Open Information Extraction with Tree Kernels. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, GA, USA, 9–14 June 2013; pp. 868–877.
4. Romero, J.; Razniewski, S. Mapping and Cleaning Open Commonsense Knowledge Bases with Generative Translation. In *The Semantic Web—ISWC 2023*; Springer: Cham, Switzerland, 2023; Volume 14265, pp. 321–337. [CrossRef]
5. Ji, S.; Pan, S.; Cambria, E.; Marttinen, P.; Yu, P.S. A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *33*, 494–514. [CrossRef] [PubMed]
6. Shi, L.; Li, S.; Yang, X.; Qi, J.; Pan, G.; Zhou, B. Semantic health knowledge graph: Semantic integration of heterogeneous medical knowledge and services. *BioMed Res. Int.* **2017**, *2017*, 2858423. [CrossRef]
7. Bizon, C.; Cox, S.; Balhoff, J.; Kebede, Y.; Wang, P.; Morton, K.; Fecho, K.; Tropsha, A. ROBOKOP KG and KGB: Integrated knowledge graphs from federated sources. *J. Chem. Inf. Model.* **2019**, *59*, 4968–4973. [CrossRef]
8. Li, L.; Wang, P.; Yan, J.; Wang, Y.; Li, S.; Jiang, J.; Sun, Z.; Tang, B.; Chang, T.H.; Wang, S.; et al. Real-world data medical knowledge graph: Construction and applications. *Artif. Intell. Med.* **2020**, *103*, 101817. [CrossRef]
9. Biswas, S.; Mitra, P.; Rao, K.S. Relation Prediction of Co-Morbid Diseases Using Knowledge Graph Completion. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2021**, *18*, 708–717. [CrossRef]
10. Du, J.; Li, X. A knowledge graph of combined drug therapies using semantic predications from biomedical literature: Algorithm development. *JMIR Med. Inform.* **2020**, *8*, e18323. [CrossRef]
11. Rindflesch, T.C.; Fiszman, M. The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text. *J. Biomed. Inform.* **2003**, *36*, 462–477.
12. Wei, C.H.; Kao, H.Y.; Lu, Z. PubTator: A web-based text mining tool for assisting biocuration. *Nucleic Acids Res.* **2013**, *41*, W518–W522. [CrossRef]
13. Fang, Y.; Wang, H.; Wang, L.; Di, R.; Song, Y. Diagnosis of COPD Based on a Knowledge Graph and Integrated Model. *IEEE Access* **2019**, *7*, 46004–46013. [CrossRef]
14. Li, R.; Yin, C.; Yang, S.; Qian, B.; Zhang, P. Marrying Medical Domain Knowledge With Deep Learning on Electronic Health Records: A Deep Visual Analytics Approach. *J. Med. Internet Res.* **2020**, *22*, e20645. [CrossRef] [PubMed]
15. Zhang, J.; Gong, J.; Barnes, L. HCNN: Heterogeneous convolutional neural networks for comorbid risk prediction with electronic health records. In Proceedings of the 2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), Philadelphia, PA, USA, 17–19 July 2017; pp. 214–221.
16. Vretinaris, A.; Lei, C.; Efthymiou, V.; Qin, X.; Özcan, F. Medical Entity Disambiguation Using Graph Neural Networks. In Proceedings of the 2021 International Conference on Management of Data, SIGMOD/PODS '21, Virtual, 20–25 June 2021; ACL: New York, NY, USA, 2021; pp. 2310–2318. [CrossRef]
17. Crichton, G.K.O. Improving Automated Literature-Based Discovery with Neural Networks: Neural Biomedical Named Entity Recognition, Link Prediction and Discovery. Ph.D. Thesis, University of Cambridge, Cambridge, UK, 2019.
18. Sheikhalishahi, S.; Miotto, R.; Dudley, J.T.; Lavelli, A.; Rinaldi, F.; Osmani, V. Natural language processing of clinical notes on chronic diseases: Systematic review. *JMIR Med. Inform.* **2019**, *7*, e12239. [CrossRef] [PubMed]
19. Yu, K.H.; Beam, A.L.; Kohane, I.S. Artificial intelligence in healthcare. *Nat. Biomed. Eng.* **2018**, *2*, 719–731. [CrossRef] [PubMed]
20. Mesquita, F.; Schmidek, J.; Barbosa, D. Effectiveness and Efficiency of Open Relation Extraction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013*; Yarowsky, D., Baldwin, T., Korhonen, A., Livescu, K., Bethard, S., Eds.; Association for Computational Linguistics: Washington, DC, USA, 2013; pp. 447–457.
21. Lin, B.Y.; Chen, X.; Chen, J.; Ren, X. KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; ACL: New York, NY, USA, 2019; pp. 2829–2839.
22. Bosselut, A.; Rashkin, H.; Sap, M.; Malaviya, C.; Celikyilmaz, A.; Choi, Y. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; ACL: New York, NY, USA, 2019; pp. 4762–4779.
23. García-Durán, A.; Dumančić, S.; Niepert, M. Learning Sequence Encoders for Temporal Knowledge Graph Completion. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 4816–4821. [CrossRef]
24. Dasgupta, S.S.; Ray, S.N.; Talukdar, P. HyTE: Hyperplane-based Temporally Aware Knowledge Graph Embedding. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 2001–2011. [CrossRef]
25. Han, X.; Cao, S.; Lv, X.; Lin, Y.; Liu, Z.; Sun, M.; Li, J. OpenKE: An Open Toolkit for Knowledge Embedding. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Brussels, Belgium, 31 October–4 November 2018; pp. 139–144. [CrossRef]

26. Guo, S.; Wang, Q.; Wang, L.; Wang, B.; Guo, L. Knowledge Graph Embedding with Iterative Guidance from Soft Rules. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 3–7 February 2018; pp. 4816–4823. [CrossRef]
27. Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; Yakhnenko, O. Translating Embeddings for Modeling Multi-relational Data. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 2787–2795.
28. Yang, B.; Yih, W.t.; He, X.; Gao, J.; Deng, L. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
29. Schlichtkrull, M.; Kipf, T.N.; Bloem, P.; van den Berg, R.; Titov, I.; Welling, M. Modeling Relational Data with Graph Convolutional Networks. In Proceedings of the The Semantic Web, Heraklion, Crete, Greece, 3–7 June 2018; pp. 593–607. [CrossRef]
30. Vashishth, S.; Sanyal, S.; Nitish, V.; Talukdar, P. Composition-based Multi-Relational Graph Convolutional Networks. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020.
31. Hwang, J.D.; Bhagavatula, C.; Le Bras, R.; Da, J.; Sakaguchi, K.; Bosselut, A.; Choi, Y. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 6384–6392.
32. Roberts, A.; Raffel, C.; Shazeer, N. How Much Knowledge Can You Pack Into the Parameters of a Language Model? In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; Association for Computational Linguistics: New York, NY, USA, 2020.
33. Zhao, Y.; Feng, H.; Gallinari, P. Embedding learning with triple trustiness on noisy knowledge graph. *Entropy* **2019**, *21*, 1083. [CrossRef]
34. Vashishth, S.; Jain, P.; Talukdar, P. Cesi: Canonicalizing open knowledge bases using embeddings and side information. In Proceedings of the 2018 World Wide Web Conference, Lyon, France, 23–27 April 2018; pp. 1317–1327.
35. Etzioni, O.; Banko, M.; Soderland, S.; Weld, D.S. Open information extraction from the web. *Commun. ACM* **2008**, *51*, 68–74. [CrossRef]
36. Fader, A.; Soderland, S.; Etzioni, O. Identifying relations for open information extraction. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Scotland, UK, 27–31 July; Association for Computational Linguistics: New York, NY, USA, 2011; pp. 1535–1545.
37. Cetto, M.; Niklaus, C.; Freitas, A.; Handschuh, S. Graphene: A Context-Preserving Open Information Extraction System. In Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, Santa Fe, NM, USA, 20–26 August 2018; ACL: New York, NY, USA, 2018; pp. 94–98.
38. Hays, D.G. Dependency Theory: A Formalism and Some Observations. *Language* **1964**, *40*, 511–525. [CrossRef]
39. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems—NeurIPS, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
40. Luong, T.; Pham, H.; Manning, C.D. Effective Approaches to Attention-based Neural Machine Translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; ACL: New York, NY, USA, 2015; pp. 1412–1421. [CrossRef]
41. Fang, T.; Wang, W.; Choi, S.; Hao, S.; Zhang, H.; Song, Y.; He, B. Benchmarking Commonsense Knowledge Base Population with an Effective Evaluation Dataset. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online, 7–11 November 2021; ACL: New York, NY, USA, 2021; pp. 8949–8964.
42. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. In Proceedings of the ICLR 2015: International Conference on Learning Representations 2015, San Diego, CA, USA, 7–9 May 2015.
43. Dutta, S.; Gautam, T.; Chakrabarti, S.; Chakraborty, T. Redesigning the Transformer Architecture with Insights from Multi-particle Dynamical Systems. *arXiv* **2021**, arXiv:2109.15142.
44. Resnik, P. Selectional Preference and Sense Disambiguation. In *Tagging Text with Lexical Semantics: Why, What, and How?*; Association for Computational Linguistics: Washington DC, USA, 1997.
45. Plank, F. Verbs and objects in semantic agreement: Minor differences between English and German that might suggest a major one. *J. Semant.* **1984**, *3*, 305–360. [CrossRef]
46. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [CrossRef]
47. Arroyo-Fernández, I.; Meza-Ruiz, I. LIPN-IIMAS at SemEval-2017 Task 1: Subword embeddings, attention recurrent neural networks and cross word alignment for semantic textual similarity. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, BC, Canada, 3–4 August 2017; ACL: New York, NY, USA, 2017, pp. 208–212.
48. Arroyo-Fernández, I.; Méndez-Cruz, C.F.; Sierra, G.; Torres-Moreno, J.M.; Sidorov, G. Unsupervised sentence representations as word information series: Revisiting TF–IDF. *Comput. Speech Lang.* **2019**, *56*, 107–129. [CrossRef]

49. Zapirain, B.; Agirre, E.; Marquez, L.; Surdeanu, M. Selectional preferences for semantic role classification. *Comput. Linguist.* **2013**, *39*, 631–663. [CrossRef]
50. Del Corro, L.; Gemulla, R. ClausIE: Clause-Based Open Information Extraction. In Proceedings of the 22nd International Conference on World Wide Web, WWW '13, Rio de Janeiro, Brazil, 13–17 May 2013; ACM: New York, NY, USA, 2013; pp. 355–366. [CrossRef]
51. Gashteovski, K.; Gemulla, R.; del Corro, L. MinIE: Minimizing Facts in Open Information Extraction. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 9–11 September 2017; ACL: New York, NY, USA, 2017; pp. 2630–2640. [CrossRef]
52. Bhardwaj, S.; Aggarwal, S.; Mausam, M. CaRB: A Crowdsourced Benchmark for Open IE. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; ACL: New York, NY, USA, 2019; pp. 6262–6267. [CrossRef]
53. Lechelle, W.; Gotti, F.; Langlais, P. WiRe57: A Fine-Grained Benchmark for Open Information Extraction. In Proceedings of the 13th Linguistic Annotation Workshop, Florence, Italy, 1 August 2019; ACL: New York, NY, USA, 2019; pp. 6–15. [CrossRef]
54. World Health Organization. *Noncommunicable Diseases Country Profiles 2018*; WHO: Geneva, Switzerland, 2018.
55. Manning, C.D.; Surdeanu, M.; Bauer, J.; Finkel, J.R.; Bethard, S.; McClosky, D. The Stanford CoreNLP natural language processing toolkit. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Baltimore, MD, USA, 23–24 June 2014; ACL: New York, NY, USA, 2014; pp. 55–60.
56. Saha, S.; Pal, H.; Mausam. Bootstrapping for Numerical Open IE. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Vancouver, BC, Canada, 30 July–4 August 2017; ACL: New York, NY, USA, 2017; pp. 317–323. [CrossRef]
57. Li, X.; Taheri, A.; Tu, L.; Gimpel, K. Commonsense knowledge base completion. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 7–12 August 2016; ACL: New York, NY, USA, 2016; pp. 1445–1455.
58. Dhingra, B.; Zaheer, M.; Balachandran, V.; Neubig, G.; Salakhutdinov, R.; Cohen, W.W. Differentiable reasoning over a virtual knowledge base. *arXiv* **2020**, arXiv:2002.10640.
59. Lakew, S.M.; Cettolo, M.; Federico, M. A Comparison of Transformer and Recurrent Neural Networks on Multilingual Neural Machine Translation. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; ACL: New York, NY, USA, 2018; pp. 641–652.
60. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [CrossRef]
61. Probst, P.; Boulesteix, A.L.; Bischl, B. Tunability: Importance of hyperparameters of machine learning algorithms. *J. Mach. Learn. Res.* **2019**, *20*, 1–32.
62. Augustin, L.; Gallus, S.; Negri, E.; La Vecchia, C. Glycemic index, glycemic load and risk of gastric cancer. *Ann. Oncol.* **2004**, *15*, 581–584. [CrossRef]
63. Lachin, J.M.; Genuth, S.; Nathan, D.M.; Zinman, B.; Rutledge, B.N.; DCCT/EDIC Research Group. Effect of glycemic exposure on the risk of microvascular complications in the diabetes control and complications trial—revisited. *Diabetes* **2008**, *57*, 995–1001. [CrossRef]
64. Chen, C.; Cai, F.; Hu, X.; Chen, W.; Chen, H. HHGN: A Hierarchical Reasoning-based Heterogeneous Graph Neural Network for Fact Verification. *Inf. Process. Manag.* **2021**, *58*, 102659. [CrossRef]
65. Le, M.N. Error propagation. Ph.D. Thesis, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands, 2021.
66. Sansford, H.; Richardson, N.; Maretic, H.P.; Saada, J.N. GraphEval: A Knowledge-Graph Based LLM Hallucination Evaluation Framework. *arXiv* **2024**, arXiv:2407.10793.
67. Baker, C.F. FrameNet: Frame Semantic Annotation in Practice. In *Handbook of Linguistic Annotation*; Ide, N., Pustejovsky, J., Eds.; Springer, Dordrecht, The Netherlands, 2017. [CrossRef]
68. Zhang, X.; Wang, Y.; Yao, J.; Sun, Y. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *arXiv* **2023**, arXiv:2311.05232.
69. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Rajani, N.; Chen, Y.; Song, Y.; et al. Survey of Hallucination in Natural Language Generation. *arXiv* **2022**, arXiv:2202.03629. [CrossRef]
70. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.t.; Edunov, S.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, BC, Canada, 6–12 December 2020; Volume 33, pp. 9459–9474.

71. Farquhar, S.; Kossen, J.; Kuhn, L.; Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature* **2024**, *630*, 625–630. [CrossRef] [PubMed]
72. Filippova, K. Controlled Hallucinations:Learning to Generate Faithfully from Noisy Data. In Proceedings of the Findings of EMNLP 2020, Online, 16–20 November 2020.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.