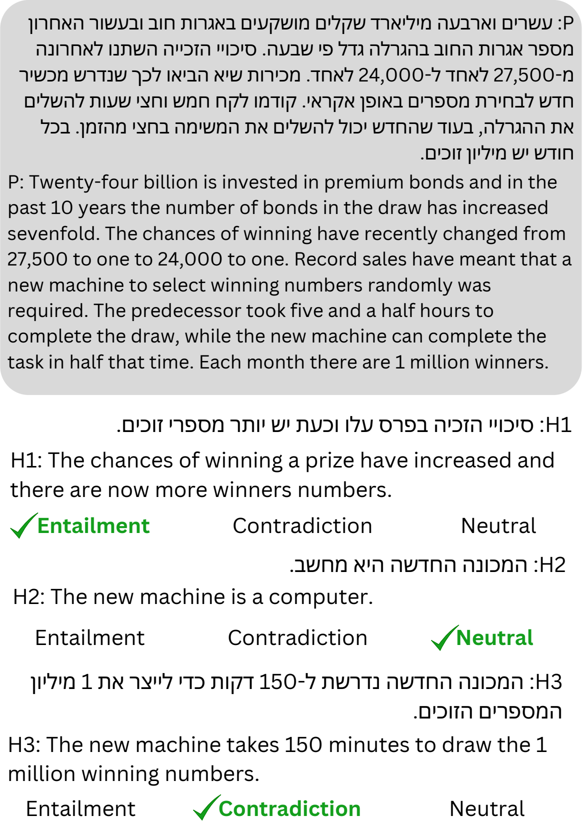
LCHAIM - Investigating Long Context Reasoning in Hebrew

Anonymous ACL submission

Abstract

Natural Language Inference (NLI) has gained significant attention recently due to its importance in understanding how machines comprehend and reason about language. While English has received tremendous interest, Morphologically Rich Languages (MRLs) like Hebrew, require more research. In this paper, we address the evaluation of Hebrew NLI models by introducing LCHAIM, a dataset designed to evaluate these models on tasks involving long premises and complex reasoning. The dataset, created by translating and validating the English ConTRoL dataset, consists of 8,325 context-hypothesis pairs that require coreferential, temporal, logical and analytical reasoning. Our experiments show the difficulty of contextual reasoning in Hebrew, as evidenced by the performance of different models. Fine-tuning the LongHero model on both the shorter premise Hebrew NLI and the LCHAIM datasets yielded a mean accuracy of 52%, that is 35% less than human performance. Similarly, Large language Models (LLMs) like Gemma-9B, Dicta-LM-2.0-7B, and GPT-4o achieved a top mean accuracy of 60.12% in few-shot setting.



P: עשרים וארבעה מיליארד שקלים מושקעים באגרות חוב ובעשור האחרון
מספר אגרות החוב בהגרלה גדל פי שבעה. סיכויי הזכייה השתנו לאחרונה
מ-27,500 לאחד ל-24,000 לאחד. מכירות שיא הביאו לכך שנדרש מכשיר
חדש לבחירת מספרים באופן אקראי. קודמו לקח חמש וחצי שעות להשלים
את ההגרלה, בעוד שהחדש יכול להשלים את המשימה בחצי מהזמן. בכל
חודש יש מיליון זוכים.

P: Twenty-four billion is invested in premium bonds and in the past 10 years the number of bonds in the draw has increased sevenfold. The chances of winning have recently changed from 27,500 to one to 24,000 to one. Record sales have meant that a new machine to select winning numbers randomly was required. The predecessor took five and a half hours to complete the draw, while the new machine can complete the task in half that time. Each month there are 1 million winners.

H1: סיכויי הזכיה בפרס עלו וכעת יש יותר מספרי זוכים.

H1: The chances of winning a prize have increased and there are now more winners numbers.

✓**Entailment** Contradiction Neutral

H2: המכונה החדשה היא מחשב.

H2: The new machine is a computer.

Entailment Contradiction ✓**Neutral**

H3: המכונה החדשה נדרשת ל-150 דקות כדי לייצר את 1 מיליון
המספרים הזוכים.

H3: The new machine takes 150 minutes to draw the 1 million winning numbers.

Entailment ✓**Contradiction** Neutral

Figure 1: A sample of NLI task from the proposed LCHAIM dataset and ConTRoL dataset.

1 Introduction

NLI, also called Textual Entailment (TE) is a key task in natural language processing (Dagan et al., 2013; Bowman et al., 2015). It involves determining whether a given hypothesis can be logically inferred from a premise. This task has shown to be very helpful in various applications, including text classification, event extraction, and summarization evaluation, (Yin et al., 2019; Sainz et al., 2022; Scirè et al., 2024). NLI has gained significant attention recently due to its importance in understanding how machines comprehend and reason about language. This is largely because almost any task can be generalized to determining entailment or contradiction in context between texts (Liu et al., 2021a).

Although many datasets have been developed to train and evaluate NLI models (Dagan et al., 2005; Giampiccolo et al., 2007; Bowman et al., 2015; Williams et al., 2018; Welleck et al., 2019), these datasets primarily focus on sentence-level examples, which do not fully capture the complexity of real-world language understanding.

To address this issue, Liu et al. (2021a) proposed ConTRoL, a long context NLI dataset that enabled investigating contextual reasoning for NLI [1]. The

[1] Other English datasets with long context NLI include (Yin et al., 2021; Koreeda and Manning, 2021; Shaham et al., 2022), but these do not specifically include complex reasoning.

example in Figure 1 demonstrates Mathematical reasoning, which requires an understanding of calculus to solve it.

However, despite large advances in NLP, research and datasets predominantly focus on English, leaving rich morphological languages like Hebrew underrepresented. On the other hand, the rich morphological structure of Hebrew presents significant challenges for Natural Language Understanding (NLU) tasks, leading for example to inconsistencies and errors in evaluation metrics (Cohen et al., 2023).

Recent studies have increased interest in Hebrew (Seker et al., 2022; Gueta et al., 2022), but the emphasis remains largely on morpho-syntactic tasks. Despite the existence of a few NLI datasets for MRLs (Klemen et al., 2024; Halat and Atlamaz, 2024; Jallad and Ghneim, 2022), to the best of our knowledge, no existing dataset specifically addresses the challenges associated with long premises that require complex reasoning. Furthermore, none of the aforementioned datasets is in Hebrew.

To address these gaps, we present in this paper a first investigation of long context reasoning in Hebrew. As part of this investigation, we compile LCHAIM - **L**ong **C**ontext **H**ebrew with **A**dvanced reasoning **I**nference **M**odel Benchmark - a Hebrew version of ConTRoL (Liu et al., 2021a). LCHAIM contains 8,325 premise-hypothesis pairs in Hebrew, labeled for contradiction, entailment, or neutral relationship between premise p and hypothesis h, and is obtained from the ConTRoL dataset via automatic translation and human validation. Then, we provide an evaluation of Hebrew Language Models for the task of NLI, using the resulting dataset. We show that the task of NLI with contextual reasoning is challenging even for larger language models, which shows the need for the construction of complex NLU benchmarks in Hebrew.

The research contribution is three-fold:

1. We present LCHAIM, a dataset for training, and validating models on complex reasoning in Hebrew.

2. We adapt, validate and document a translation methodology for translating datasets from English to Hebrew.

3. We perform detailed experiments to evaluate Hebrew models on the task of long-premise NLI with complex reasoning.

2 LCHAIM Dataset

The LCHAIM dataset is a translated version of ConTRoL (Liu et al., 2021a). It is compiled from publicly available online verbal reasoning tests, including those used in police initial recruitment exams, medical college admissions, and university clinical aptitude tests, as well as corporate verbal aptitude assessments. These tests are structurally similar to NLI tasks, presenting a premise and a hypothesis with three answer choices: true, false, or cannot say. These choices correspond to the NLI labels ENTAILMENT, CONTRADICTION, and NEUTRAL.

The data format of LCHAIM aligns with existing NLI benchmarks (Bowman et al., 2015; Williams et al., 2018). A notable difference from existing datasets is that LCHAIM features longer premises, often spanning one or more passages. Additionally, each premise is paired with three or more hypotheses.

The verbal reasoning tests require candidates to comprehend meaning, evaluate logical strength, make valid inferences, and identify appropriate conclusions. The passages cover various topics, including current affairs, business, science, the environment, economics, history, meteorology, health, and education. These questions are of high quality and are used in rigorous assessments, reflecting a high level of difficulty. After removing duplicates, the LCHAIM dataset consists of 8,325 context-hypothesis pairs. Lexical overlap analysis, calculated using Jaccard Similarity between premises and hypotheses in each class, shows only 4.87% overlap for ENTAILMENT pairs and 5.49% for CONTRADICTION pairs, indicating that the dataset presents significant challenges for simple lexical matching techniques (Liu et al., 2021a).

Inferring from the original ConTRoL dataset, and with the hypothesis that translation does not affect the reasoning types present in the dataset at large. LCHAIM contains various reasoning tasks such as Coreferential Reasoning (Ye et al., 2020), Verbal Logical Reasoning (Liu et al., 2021b), Temporal and Mathematical Reasoning (Nakhimovsky, 1987), Information Integration over Paragraphs (Welbl et al., 2018), and Analytical Reasoning (Williams et al., 2019).

2.1 Translation

The LCHAIM dataset was created via translation of the previously mentioned ConTRoL dataset, a large

NLI dataset with especially long premises that aims to address contextual reasoning. To ensure optimal translation we employ two translation models, Amazon Translate[2], a neural network-based machine translation service, and DictaLM Translation, a Hebrew-LLM-based translation service (Shmidman et al., 2024). We translate the entire ConTRoL dataset using both translation models using the API of both services.

2.2 Validation

To determine which translation is superior for this task, we perform a manual evaluation of the translation.

We employed a stratified random sampling approach. From our dataset of 8,325 samples, we randomly selected 200 samples, maintaining the proportional representation of the original train, test, and development splits.

Two out of four human annotators (two authors of the paper) independently reviewed and ranked each translated sample based on three ordinal scale parameters and one binary parameter:

1. Translation Accuracy: The degree of precision in translating from English to Hebrew (Popović et al., 2006).

2. Fluency: The naturalness and coherence of the translated premise and hypothesis (Graham et al., 2013).

3. Cultural Fit: The extent to which the translation preserves the cultural nuances and appropriateness of the original English text in Hebrew (Nida, 1964).

4. Label Validity: Binary, whether the label is still valid after the translation of the premise and hypothesis.

Each ordinal parameter was scored on a scale of 1 to 5, with 5 representing the highest quality and 1 the lowest. Additionally, the annotators assessed whether the original label remained valid after translation, ensuring the logical relationship between the premise and hypothesis is reserved in the target language.

Following this validation process, we calculated the average annotators' score for each of the three parameters (Translation Accuracy, Fluency, and Cultural Fit). We also computed the percentage of cases where the original label was maintained after translation. These results are presented in Table 1.

Amazon Translate was selected as the better translation in the context of this dataset, with a mean accuracy, fluency, and cultural fit of 85.57%, 82.53% ,74.49% respectively. The percentage of samples with original labels remaining valid after translation was 97.8% for Amazon Translate and 94% for DictaLM Translation. This evaluation methodology allowed us to comprehensively assess the quality and reliability of both automatic translation models in the context of our specific task, with quantitative measures to support our choice.

Additionally, we conduct an automatic validation of the translation quality using back-translation. Specifically, we re-translate both premises and hypotheses back to English using the same translation service and compute sentence similarity between the original English text and its back-translation. For this evaluation, we used MiniLM[3], which yielded an average sentence similarity of 0.9537 ± 0.0479 for premises, with only 22 out of unique 1928 unique instances scoring below 0.75, following (Lin et al., 2021). For hypotheses, the average sentence similarity is 0.9331 ± 0.0886, with 103 out of unique 1928 instances scoring below the 0.75 threshold. Based on these results, we remove samples with a sentence similarity < 0.75 from the dataset to ensure dataset quality.

To further verify that the back-translated dataset preserves semantics, we extract embeddings from both the original dataset (ConTRoL) and its back-translated version using a Longformer (Beltagy et al., 2020) model. We then train a Multi-Layer Perceptron (MLP) classifier solely on the ConTRoL train set embeddings and use it to make predictions on both the original and LCHAIM-back-translated test sets. When comparing test results, we find that the classifier assigns the same label (entailment, contradiction, or neutral) to 92.11% of the test samples in both datasets. This high agreement indicates that the translation process largely preserves the essential meaning of the dataset.

3 Experiments

3.1 Hebrew Pre-trained Language Models

We implement and evaluate SOTA pre-trained language models in Hebrew to demonstrate their

[2] https://aws.amazon.com/translate/

[3] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Model	Accuracy	Fluency	Cult. Fit	Label Validity
Amazon Translate	4.279 ± 0.075	4.127 ± 0.0495	3.775 ± 0.09	4.890 ± 0.01
DICTA-LM-2.0	4.057 ± 0.057	4.096 ± 0.031	3.927 ± 0.148	4.700 ± 0.190

Table 1: Translation evaluation metrics for Amazon Translate and DICTA-LM-2.0. Scores range from 1 to 5 (best). 95% confidence interval for the mean absolute error for inter-rater agreement in manual validation is also reported. Bold indicates the best rated translation.

NLU capabilities, proxied by their performance on LCHAIM.

AlephBERT AlephBERT (Seker et al., 2022) is a Hebrew version of the classic BERT (Devlin et al., 2019) pre-trained language model with a large vocabulary. It was trained on three datasets: the Hebrew portion of the OSCAR corpus-cleaned Twitter texts (Ortiz Su'arez et al., 2020), and the entire Hebrew Wikipedia. Both variants utilize wordpieces with a vocabulary size of 52,000. AlephBERT was trained only using the masked token prediction objective, excluding next-sentence prediction.

LongHero LongHero (Shalumov and Haskey, 2023) is a variant of the HeRo model (Shalumov and Haskey, 2023) which is a Hebrew language model based on the RoBERTa architecture (Liu et al., 2019). It is designed to handle long sequences more effectively and uses a BPE tokenizer with a vocabulary size of 50,265 tokens, trained on the HeDC4 dataset (Shalumov and Haskey, 2023).

We implement this experiment following the original ConTRoL paper implementation. Given a premise p and a hypothesis h, we concatenate them into a new sequence: $[\texttt{CLS}] + p + [\texttt{SEP}] + h + [\texttt{SEP}]$, where $[\texttt{CLS}]$ is the classification token and $[\texttt{SEP}]$ is the separator token. We encode the sequence with the pre-trained model, and feed the hidden representation of the $[\texttt{CLS}]$ token from the final layer to the Multi-Layer Perceptron (MLP) with a softmax layer for classification. The MLP has three hidden layers followed by an output layer. The first performs a linear transformation from the input size to 300 units, followed by ReLU activation, layer normalization, and dropout with a rate of 0.3. The second layer transforms to 100 units, and the third to 50 units. The final output layer maps the 50 units to 3 output classes for classification. Each hidden layer includes ReLU activation and layer normalization.

For AlephBERT, we use the proposed methodology by Devlin et al. (2019). We adhere to the original implementation details, using the same dataset split into training, development, and test sets with an 8:1:1 ratio (6,692:837:836 samples for training, validation and test respectively) and training all models for 10 epochs (see Appendix A for hyperparameter optimization details). The maximum sequence length is set to 512 tokens for AlephBERT and 4,096 for LongHero.

To further investigate the performance of these models in different settings, we employ fine-tuning in 3 ways. Firstly, we fine-tune both AlephBERT and LongHero using the HebNLI (HebArabNlp-Project, 2024) dataset, a translated version of the original SNLI dataset (Bowman et al., 2015). This is done to investigate the effect of fine-tuning using a short-premise NLI dataset for the task of long-premise NLI in LCHAIM. Second, we perform further fine-tuning of the same models using the train set (80% of samples) of the LCHAIM dataset. We report Accuracy over the test set as the main metric, and precision (P), recall (R), and F1-Score (F1) for each class.

3.2 In-Context Learning with LLMs

We further evaluate three SOTA Hebrew and multilingual LLMs for the LCHAIM NLI task: (1) **Dicta-LM-2.0-7B** (Shmidman et al., 2024), (2) **Gemma-9B** (Team et al., 2024), and (3) **GPT-4o** (Hurst et al., 2024) For models description, see Appendix B. Inference with LLMs was performed using the Huggingface Transformers ecosystem (Wolf, 2019) and Azure's OpenAI Service (Microsoft, 2024), by parsing the LLM's response. Each model returned a single-letter label: 'e' for entailment, 'c' for contradiction, and 'n' for neutral. We ensured that every response adhered to this format, validating the outputs accordingly. results are reported over the test set only. During the prompt engineering process, several methodologies were investigated to enhance the model's performance. Initially, we employed a zero-shot approach, wherein only free-text instructions were provided to the model to guide the task. However, this method proved insufficient, as it was overly simplistic for the models to accurately predict the specific labels, resulting in incorrect outputs.

We then implemented a prompt with detailed instructions regarding the structure of the desired output (see Appendix C). This adjustment led to an improvement in the model's performance. To further refine the approach, we iteratively modified the prompt in the zero-shot setting, randomly sampling and evaluating 30 instances after each iteration. If performance improved, we retained the new prompt; otherwise, we reverted to the previous version. This process was repeated over seven iterations. Finally, we introduced examples, known as the "few-shot" (Brown, 2020) approach or in-context learning, which involved providing the model with several instances of NLI tasks along with the correct labels.

Additional Experiments To obtain a human performance baseline, we perform a manual evaluation, where 4 of the authors, Hebrew native speakers, assessed 200 random samples from the test set of the LCHAIM dataset.

To further investigate the complexity of the various types of reasoning, and the increased context length available in LCHAIM, we test AlephBERT, LongHero, and Gemma-9B using the shorter-premise HebNLI dataset, a translated version of the original Stanford NLI (SNLI, Bowman et al., 2015), for comparison. We initially ran models without fine-tuning, to highlight the impact of fine-tuning and compare baseline performance between models.

4 Results

Table 2 presents the results for the models tested. AlephBERT, achieved an overall mean accuracy of 39.5%, with an F1-Score of 36. However, the model showed relatively uneven calibration, consistently predicting the entailment class in the test set, after accounting for overfitting through class balance, vanishing gradient, and varying model complexity, we assume this behavior indicates the difficulty of the task at hand. Fine-tuning AlephBERT using the HebNLI dataset achieved 37.5% mean accuracy and 35.42 F1 Score resulting in a -2%, and -1.5 absolute difference respectively, compared to the non fine-tuned model.

Fine-tuning only using LCHAIM, without the HebNLI step resulted in a mean accuracy of 38.4% and an F1 Score of 32.4, which indicates an absolute difference of -1%, and -3 respectively. Continuous fine-tuning on both HebNLI and LCHAIM resulted in a mean accuracy of 42.6% and an F1 Score of 41.3 or an overall 3%, and 5 absolute increase respectively. This means for AlephBERT the best setting is fine-tuning on both the short premise HebNLI and the long premise LCHAIM dataset, these settings also showed better calibration than the original pre-trained version, with relatively balanced class prediction.

LongHero, which has a long enough context to be able to accommodate the long premises, was able to achieve a mean accuracy and an F1-Score of 34.5%, and 33.7 respectively without fine-tuning. The model displayed some imbalance towards the entailment class. Fine-tuning LongHero using the shorter premise HebNLI resulted in an accuracy of 32% and an F1 Score of 25, or an absolute difference of -2% and -8 compared to the non-fine-tuned version. Fine-tuning only using the LCHAIM training set resulted in an accuracy of 41% and an F1 Score of 34 or an absolute difference of 7% and 1 respectively. Furthermore, fine-tuning LongHero on both HebNLI, and then on LCHAIM, resulted in the best overall performance of 52% mean accuracy and F1 Score, an absolute performance difference of 18% and 19 respectively. This can also be viewed as a relative increase of an impressive 35% increase in mean accuracy and F1 Score, which indicates a significant increase in performance. This model also showed the overall best calibration between the 3 predicted classes. The best model for the task was a fine-tuned LongHero using both HebNLI and LCHAIM.

For the LCHAIM test set, Gemma-9B achieved 56.15% accuracy in zero-shot, and 50.93%, 49.94%, and 48.45% for 1, 2, and 3-shot prompting from the HebNLI dataset. With random sampling from the LCHAIM dataset, it achieved 56.15%, 55.65%, 49.57%, and 49.32% for zero, 1, 2, and 3-shot, respectively. Dicta-LM-2.0-7B scored 39% in zero-shot, with a small 1.5% increase in accuracy for few-shot (40.45%, 40.99%, and 40.62% for 1, 2, and 3-shot). For shots sampled from LCHAIM, it achieved 42.36%, 40%, and 41.37% for 1, 2, and 3-shot. GPT-4o performed the best with 57.68%, 58.37%, 60.12%, and 57.21% for zero, 1, 2, and 3-shot, respectively. Results are shown in Table 3.

For the manual evaluation, human performance is displayed in Table 2. The testers were able to achieve a mean accuracy of 84.97% and an F1 Score of 84.26, a significant difference of 33 abso-

Model	Acc.	F1	Entailment			Contradiction			Neutral		
			P	R	F1	P	R	F1	P	R	F1
Human (English)	87.06	93.15	94.83	95.65	95.24	93.33	91.21	92.26	93.02	91.91	92.95
Human (Hebrew)	84.97	84.26	77.05	87.04	81.74	91.23	78.79	84.55	85.33	87.67	86.49
AlephBERT	39.50	35.94	42.47	68.19	52.34	31.38	24.38	27.44	39.13	15.25	21.95
AlephBERT-HebNLI	37.51	35.42	42.43	56.57	48.49	28.92	14.46	19.28	33.06	34.74	33.88
AlephBERT-LCHAIM	38.38	32.39	42.61	60.85	50.12	32.82	44.21	37.67	25.00	0.01	0.02
AlephBERT-ALL	42.60	41.32	46.26	56.88	51.02	36.45	43.38	39.62	45.21	22.03	29.63
LongHero	34.53	33.74	39.42	46.17	42.53	28.7	33.47	30.91	32.85	19.49	24.46
LongHero-HebNLI	31.55	24.92	50.77	10.09	16.84	34.48	12.40	18.24	29.25	**80.93**	42.97
LongHero-LCHAIM	41.49	34.68	41.94	**81.96**	55.49	44.12	12.40	19.35	36.73	15.25	21.56
LongHero-ALL	**52.17**	**52.20**	**59.74**	57.19	**58.44**	**47.56**	**44.22**	**45.82**	**47.19**	53.39	**50.10**

Table 2: Performance results for different models over the LCHAIM test set. Metrics include overall accuracy, F1 score, precision (P), recall (R), and F1 scores for each class: Entailment, Contradiction, and Neutral. -HebNLI indicates fine-tuning using HebNLI, -LCHAIM indicates fine-tuning using LCHAIM, and -ALL indicates sequential fine-tuning on HebNLI and then LCHAIM. Human (English) and Human (Hebrew) indicate performance of four educated testees of 300, and 200 samples from ConTRoL and LCHAIM respectively. Bold indicates the best performing model.

Model	Shots' Dataset	0-shot	1-shot	2-shot	3-shot
Gemma-9B	HebNLI LCHAIM	56.15%	50.93% 55.65%	49.94% 49.57%	48.45% 49.32%
Dicta-LM-2.0-7B	HebNLI LCHAIM	39.38 %	40.45% 42.36%	40.99% 40.00%	40.62% 41.37%
GPT-4o	HebNLI LCHAIM	**57.68** %	54.13% **58.37%**	52.57% **60.12**%	54.41% **57.21%**

Table 3: Mean accuracy of LLMs Gemma-9B and Dicta-LM-2.0-7B over the LCHAIM test set, under 0-shot, 1-shot, 2-shot, and 3-shot settings. The 'Shots' Dataset column indicates the dataset from which few-shot examples were sampled. Note that for the zero-shot setting, only a single accuracy score is reported, as no examples are provided for in-context learning. Best performance model and setting in bold. GPT-4o scored the best result, 60.12% for the 2-shot setting.

lute accuracy and F1 Score compared to the best LongHero model. Human performance also show a 27.29% higher mean accuracy compared to the best LLM, GPT-4o.

Result for the shorter HebNLI are presented in Table 4. The base AlephBERT model achieved a mean accuracy and F1 Score of 53.50% and 53.42% respectively without fine-tuning. While the LongHero base model achieved a 47.17% and 46.74% mean accuracy and F1 Score respectively. Fine-tuning using the HebNLI datasets resulted in a 15% and 27% absolute increase in accuracy for AlephBERT and LongHero respectively. However, fine-tuning using the LCHAIM dataset, resulted in a 17% and 10% absolute **decrease** for both models respectively, showing that fine-tuning using LCHAIM hinders performance for the HebNLI task.

Fine-tuning using first the LCHAIM and then HebNLI dataset resulted in a mean accuracy and F1 Score of of 78.28% and 78.24% for AlephBERT, offering similar results to fine-tuning only with HebNLI . The same approach resulted in a mean accuracy and F1 Score of 83.93% and 83.78% respectively for the LongHero model, a 5.5% increase in performance compared to fine-tuning only with HebNLI.

The best model for the HebNLI test set (LongHero-ALL), achieved a 31.76% absolute higher mean accuracy compared to the best model for the LCHAIM test set (also LongHero). Both models showed good calibration over both test sets. For comparison, SOTA performance for HebNLI is reported in the Hebrew LLM leaderboard, with a mean accuracy of 95.48% for Qwen2.5-72B.[4]

[4] https://huggingface.co/spaces/

Model	Acc.	F1	Entailment			Contradiction			Neutral		
			P	R	F1	P	R	F1	P	R	F1
AlephBERT	53.50	53.42	56.14	55.22	55.31	56.11	59.38	57.84	46.23	45.57	45.40
AlephBERT-HebNLI	78.73	78.68	78.21	83.88	80.95	80.81	78.17	79.47	77.01	73.62	75.28
AlephBERT-LCHAIM	36.42	31.02	36.68	61.18	45.86	36.08	41.36	38.54	36.00	0.03	0.06
AlephBERT-ALL	78.28	78.24	77.71	82.56	80.01	80.39	78.82	79.60	76.54	72.89	74.67
LongHero	47.17	46.74	48.22	58.22	52.75	50.51	47.55	49.99	41.22	34.43	37.52
LongHero-HebNLI	74.52	73.86	76.68	81.13	81.30	72.36	74.46	76.97	72.80	62.57	67.25
LongHero-LCHAIM	37.30	31.46	37.22	79.80	51.12	38.77	14.10	21.66	32.06	15.04	21.60
LongHero-ALL	**83.93**	**83.78**	**87.41**	**86.84**	**87.13**	**84.19**	**85.01**	**84.60**	**79.78**	**79.48**	**79.63**

Table 4: Performance results for different models using the shorter-premise HebNLI test set. Metrics include overall accuracy, F1 score, precision (P), recall (R), and F1 scores for each class: Entailment, Contradiction, and Neutral. -HebNLI indicates fine-tuning using HebNLI, -LCHAIM indicates fine-tuning using LCHAIM, and -ALL indicates sequential fine-tuning with LCHAIM and then HebNLI. Numbers in bold indicate the best performing model setting.

5 Error Analysis

5.1 Morphological Richness and Errors

To explore the relationship between morphological richness and model performance, we compute the Type-to-Token Ratio (TTR) of the test set samples. TTR is calculated as the average between two ratios. The (1) ratio of distinct parts-of-speech (POS) types to the total number of POS in the text, and (2) the ratio of distinct lemmas to the total number of lemmas in the text. TTR is used to measure lexical diversity, here we follow (Kettunen, 2014) to measure morphological diversity. We hypothesize that morphological complexity in Hebrew may affect model accuracy due to challenges in tokenization, syntactic parsing, and semantic disambiguation.

We calculated TTR values for both the premise and hypothesis combined in each dataset sample and correlated them with the model's accuracy. This analysis focuses on the best-performing models, **LongHero-ALL**, and **GPT-4o** results are shown in Figure 2. Our results show that higher TTR values, indicating greater morphological diversity, are associated with more prediction errors, supporting our hypothesis.

These findings might indicate the need to account for morphological complexity when designing models for languages like Hebrew.

5.2 Reasoning Types and Errors

We hypothesise LCHAIM is a difficult NLI task due to the complex and various reasoning types it contains, unlike other Hebrew NLI datasets. To investigate the relationship between model performance and reasoning, we report the average model accuracy for each of four main reasoning types: (1) Temporal, (2) Corereferential, (3) Logical , and (4) Analytical. We rely on previous work (Liu et al., 2021a) to determine the types of reasoning. Our analysis shows a mean accuracy of 49.16%, 64.00%, 52.27%, and 53.74% for Temporal, Coreferential, Logical, and Analytical categories, respectively. This analysis also pertains to the best performing LongHero-ALL model, where each sample can contain one or more types of reasoning.

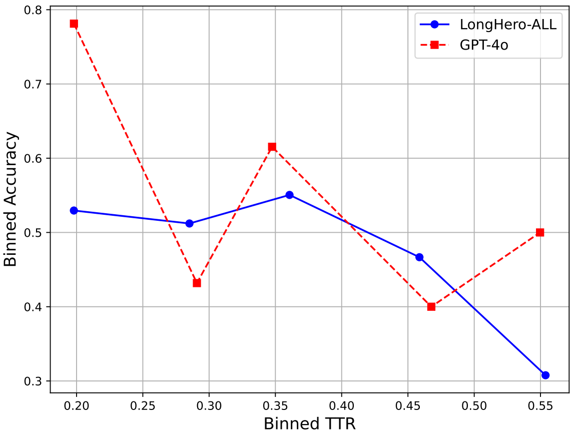


Figure 2: Mean accuracy for the best performing, fine-tuned, LongHero-ALL model, across different TTR bins, over the LCHAIM test set.

6 Discussion

The results of our evaluation of the LCHAIM dataset shed light on the challenges and performance of models dealing with long-premise NLI tasks in Hebrew. To the best of our knowledge, LCHAIM is the first long premise, complex reasoning, NLI dataset available in Hebrew.

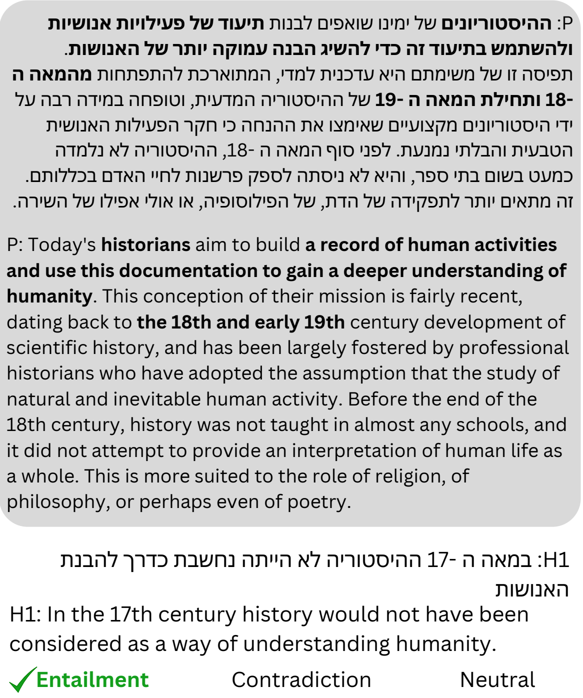
hebrew-llm-leaderboard/leaderboard

Model Performance The AlephBERT model struggled with long-premise NLI tasks. It achieved a mean accuracy of 39.50% and an F1-Score of 35.94, and tends to over-predict entailment. Fine-tuning AlephBERT on the HebNLI dataset, and on the LCHAIM train set did not improve it's performance over the test set. However, further fine-tuning the model with LCHAIM after the initial HebNLI fine-tuning showed improvements. This resulted in a mean accuracy and F1-Score of 42.60% and 41.32 respectively. This suggests the shorter, more simple task of HebNLI does not contribute towards the harder, longer-premise LCHAIM test set performance on its own, but can provide a foundation for further fine-tuning. Therefore, including the long-premise data from LCHAIM is essential for better performance in the longer premise tasks. With that said, performance was still considerably lower then the 84.97% mean accuracy human performance.

LongHero, which is designed to handle longer contexts, achieved a mean accuracy and F1-Score of 34.53% and 33.74 respectively. This model had also calibration issues similar to AlephBERT. It is worth noting the English equivalent models (BERT-base, and Longformer) did not experience these calibration issues when considering the English version of this dataset (Liu et al., 2021a). The best results were seen with the fine-tuned LongHero using both HebNLI and then LCHAIM, with a mean accuracy and F1 Score of 52.17% and 52.20 respectively. For comparison, performance over the shorter HebNLI dataset were significantly higher than the performance over the LCHAIM test set considering the sequentially fine-tuned AlephBERT and LongHero. These models showed a 25-30% higher mean accuracy and F1 Score considering the HebNLI test set. With caution, we hypothesise this is an indicator of the complexity of the reasoning in the LCHAIM dataset.

GPT-4o performed the best with a mean accuracy of 60.12% in 2-shot learning, but still did not significantly surpassed the smaller Gemma-9B. Both Gemma-9B and Dicta-LM-2.0-7B showed no improvement with few-shot learning compared to zero-shot learning.

Complex Reasoning and Morphological Richness We hypothesize that the LCHAIM dataset is a long reach to Hebrew models because, unlike previous NLI benchmarks, it contains tasks that require complex reasoning (for example temporal reasoning shown in Figure 3). Furthermore, we show that morphological richness of samples in the test set is inversely correlated with performance. This finding aligns with our hypothesis, and previous research claiming MRLs like Hebrew, are more difficult for language models than non MRLs like English.



P: **ההיסטוריונים** של ימינו שואפים לבנות **תיעוד של פעילויות אנושיות ולהשתמש בתיעוד זה כדי להשיג הבנה עמוקה יותר של האנושות**. תפיסה זו של משימתם היא עדכנית למדי, המתוארכת להתפתחות **מהמאה ה -18 ותחילת המאה ה -19** של ההיסטוריה המדעית, וטופחה במידה רבה על ידי היסטוריונים מקצועיים שאימצו את ההנחה כי חקר הפעילות האנושית הטבעית והבלתי נמנעת. לפני סוף המאה ה -18, ההיסטוריה לא נלמדה כמעט בשום בתי ספר, והיא לא ניסתה לספק פרשנות לחיי האדם בכללותם. זה מתאים יותר לתפקידה של הדת, של הפילוסופיה, או אולי אפילו של השירה.

P: Today's **historians** aim to build **a record of human activities and use this documentation to gain a deeper understanding of humanity**. This conception of their mission is fairly recent, dating back to **the 18th and early 19th** century development of scientific history, and has been largely fostered by professional historians who have adopted the assumption that the study of natural and inevitable human activity. Before the end of the 18th century, history was not taught in almost any schools, and it did not attempt to provide an interpretation of human life as a whole. This is more suited to the role of religion, of philosophy, or perhaps even of poetry.

H1: במאה ה -17 ההיסטוריה לא הייתה נחשבת כדרך להבנת האנושות

H1: In the 17th century history would not have been considered as a way of understanding humanity.

✓**Entailment** Contradiction Neutral

Figure 3: Sample from the LCHAIM dataset which requires temporal reasoning to solve.

Like ConTRoL, Hebrew models struggle to achieve human performance with LCHAIM. Unlike ConTRoL, Hebrew models show bad calibration, which could indicate specific complexities in Hebrew.

7 Conclusion

This study introduced the LCHAIM dataset, a benchmark for long-premise NLI task in Hebrew, that focuses on complex reasoning types like coreferential, temporal, and analytical reasoning. Our experiments show that current state-of-the-art models perform far below human levels. The LCHAIM dataset provides a tough challenge for future research, underscoring the need for better models and methods. By making LCHAIM and our code available, we hope to encourage further exploration and advancements in Hebrew NLU research.

Limitations

While this study advances one's knowledge about Hebrew NLI, several limitations should be acknowledged. A more granular evaluation could offer deeper insights into how models handle specific reasoning and morphological challenges in Hebrew NLI tasks. Another limitation is the fact the dataset was only evaluated manually by randomly sampling from it. Future work should dive deeper into challenges posed by advanced reasoning and morphological richness in other ways.

References

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Amir Cohen, Hilla Merhav-Fine, Yoav Goldberg, and Reut Tsarfaty. 2023. Heq: a large and diverse hebrew reading comprehension benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13693–13705.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzoto. 2013. *Recognizing Textual Entailment: Models and Applications*. Morgan and Claypool.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and William B Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pages 1–9.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41.

Eylon Gueta, Avi Shmidman, Shaltiel Shmidman, Cheyn Shmuel Shmidman, Joshua Guedalia, Moshe Koppel, Dan Bareket, Amit Seker, and Reut Tsarfaty. 2022. Large pre-trained models with extra-large vocabularies: A contrastive analysis of hebrew bert models and a new one to outperform them all. *arXiv preprint arXiv:2211.15199*.

Mustafa Halat and Ümit Atlamaz. 2024. Implicatr: A granular dataset for natural language inference and pragmatic reasoning in turkish. In *Proceedings of the First Workshop on Natural Language Processing for Turkic Languages (SIGTURK 2024)*, pages 29–41.

HebArabNlpProject. 2024. Hebrew natural language inference dataset. https://huggingface.co/datasets/HebArabNlpProject/HebNLI. Accessed: 2024-10-05.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Khloud Al Jallad and Nada Ghneim. 2022. Arnli: Arabic natural language inference for entailment and contradiction detection. *arXiv preprint arXiv:2209.13953*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

K. Kettunen. 2014. Can type-token ratio be used to show morphological complexity of languages? *Journal of Quantitative Linguistics*, 21(3):223–245.

Matej Klemen, Aleš Žagar, Jaka Čibej, and Marko Robnik-Šikonja. 2024. Si-nli: A slovene natural language inference dataset and its evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14859–14870.

Yuta Koreeda and Christopher D Manning. 2021. Contractnli: A dataset for document-level natural language inference for contracts. *arXiv preprint arXiv:2110.01799*.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021. Common sense beyond english: Evaluating and improving multilingual language models for commonsense reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1274–1287.

Hanmeng Liu, Leyang Cui, Jian Liu, and Yue Zhang. 2021a. Natural language inference in context-investigating contextual reasoning over long texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 15, pages 13388–13396.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2021b. Logiqa: a challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3622–3628.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Microsoft. 2024. Azure openai service. https://azure.microsoft.com/en-us/products/cognitive-services/openai-service. Accessed: 2025-02-05.

Alexander Nakhimovsky. 1987. Temporal reasoning in natural language understanding: The temporal structure of the narrative. In *Third Conference of the European Chapter of the Association for Computational Linguistics*.

Eugene Albert Nida. 1964. *Toward a science of translating: with special reference to principles and procedures involved in Bible translating*. Brill Archive.

Pedro Javier Ortiz Su'arez, Laurent Romary, and Benoit Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Maja Popović, Adrià de Gispert, Deepak Gupta, Patrik Lambert, Hermann Ney, José B Mariño, Marcello Federico, and Rafael E Banchs. 2006. Morpho-syntactic information for automatic error analysis of statistical machine translation output. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 1–6.

Oscar Sainz, Itziar Gonzalez-Dios, Oier Lopez de Lacalle, Bonan Min, and Eneko Agirre. 2022. Textual entailment for event argument extraction: Zero-and few-shot with multi-source learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2439–2455.

Alessandro Scirè, Karim Ghonim, and Roberto Navigli. 2024. Fenice: Factuality evaluation of summarization based on natural language inference and claim extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Amit Seker, Elron Bandel, Dan Bareket, Idan Brusilovsky, Refael Greenfeld, and Reut Tsarfaty. 2022. Alephbert: Language model pre-training and evaluation from sub-word to sentence level. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46–56.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, et al. 2022. Scrolls: Standardized comparison over long language sequences. *arXiv preprint arXiv:2201.03533*.

Vitaly Shalumov and Harel Haskey. 2023. Hero: Roberta and longformer hebrew language models. *arXiv preprint arXiv:2304.11077*.

Noam Shazeer. 2019. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*.

Noam Shazeer. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.

Shaltiel Shmidman, Avi Shmidman, Amir DN Cohen, and Moshe Koppel. 2024. Adapting llms to hebrew: Unveiling dictalm 2.0 with enhanced vocabulary and instruction capabilities. *arXiv preprint arXiv:2407.07080*.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAAfCL-HLT*, pages 1112–1122.

Chad C Williams, Mitchel Kappen, Cameron D Hassall, Bruce Wright, and Olave E Krigolson. 2019. Thinking theta and alpha: Mechanisms of intuitive and analytical reasoning. *NeuroImage*, 189:574–580.

T Wolf. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Deming Ye, Yankai Lin, Jiaju Du, Zhenghao Liu, Peng Li, Maosong Sun, and Zhiyuan Liu. 2020. Coreferential reasoning learning for language representation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7170–7186.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. 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)*, pages 3914–3923.

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. Docnli: A large-scale dataset for document-level natural language inference. *arXiv preprint arXiv:2106.09449*.

Biao Zhang and Rico Sennrich. 2019. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32.

A Hyperparameters Optimization Details

Hyperparameters were optimized through an extensive grid search across a variety of configurations to achieve optimal model performance. Batch sizes of 8, 16, and 32 were tested, while learning rates ranging from 1e^{-5} to 5e^{-5} (specifically 1e^{-5}, 2e^{-5}, 3e^{-5}, 4e^{-5}, 5e^{-5}) were explored to determine the optimal convergence rate. Gradient accumulation steps of 1, 2, and 4 were also evaluated. These combinations enabled us to systematically test different hyperparameter settings and identify the best configuration for each model.

B Large Language Models (LLMs) Details

Dicta-LM-2.0-7B Dicta LM 2.0 (Shmidman et al., 2024) is a Hebrew LLM with a custom tokenizer, continuously pre-trained from Mistral-7B-v0.1 (Jiang et al., 2023) and later fine-tuned. The pertaining is done using a Hebrew corpus, and further supervised fine tuning is performed using custom Hebrew datasets.

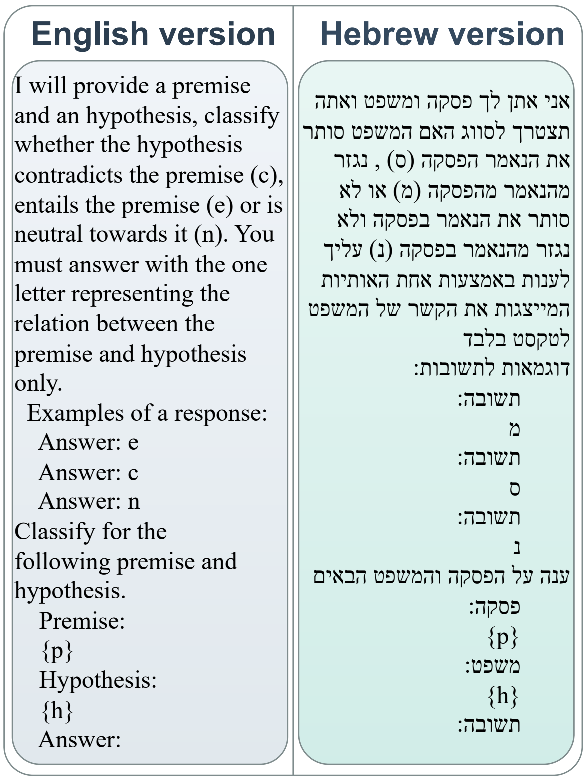
Gemma-9B The Gemma-9B model (Team et al., 2024) is a multilingual transformer decoder trained on an 8192-token context. It employs Multi-Query Attention (Shazeer, 2019) for efficiency, Rotary Positional Embeddings (RoPE) (Su et al., 2024) for compact representation, and GeGLU (Shazeer, 2020) with RMSNorm (Zhang and Sennrich, 2019) for stable training.

GPT-4o GPT-4o is a state-of-the-art multimodal LLM developed by OpenAI (Hurst et al., 2024). This model is closed-source, and benefits from extensive pre-training.

C Prompt Structure and Classification Guidelines

We provide the models with a Hebrew prompt for classifying the relationship between a premise and a hypothesis (Figure 4). The English translation mirrors the Hebrew prompt.

The are prompted to classify the premise-hypothesis relationship in a single letter as c (Contradiction): Hypothesis contradicts the premise. e (Entailment): Hypothesis follows from the premise. n (Neutral): Hypothesis neither contradicts nor follows.



English version

I will provide a premise and an hypothesis, classify whether the hypothesis contradicts the premise (c), entails the premise (e) or is neutral towards it (n). You must answer with the one letter representing the relation between the premise and hypothesis only.
Examples of a response:
Answer: e
Answer: c
Answer: n
Classify for the following premise and hypothesis.
Premise:
{p}
Hypothesis:
{h}
Answer:

Hebrew version

אני אתן לך פסקה ומשפט ואתה
תצטרך לסווג האם המשפט סותר
את הנאמר הפסקה (ס) , נגזר
מהנאמר מהפסקה (מ) או לא
סותר את הנאמר בפסקה ולא
נגזר מהנאמר בפסקה (נ) עליך
לענות באמצעות אחת האותיות
המייצגות את הקשר של המשפט
לטקסט בלבד
דוגמאות לתשובות:
תשובה:
מ
תשובה:
ס
תשובה:
נ
ענה על הפסקה והמשפט הבאים
פסקה:
{p}
משפט:
{h}
תשובה:

Figure 4: Prompt used for the NLI task. Left side shows the English translation, right side shows the original Hebrew. The prompt instructs the model to classify the relationship between a given premise and hypothesis as either contradiction (c), entailment (e), or neutral (n), using a single letter for the response.