HeSum: a Novel Dataset for Abstractive Text Summarization in Hebrew

Anonymous ACL submission

Abstract

While large language models (LLMs) excel in various natural language tasks in English, their performance in low-resource languages like Hebrew, especially for generative tasks such as abstractive summarization, remains unclear. The high morphological richness in Hebrew adds further challenges due to the ambiguity in sentence comprehension and the complexities in meaning construction. In this paper, we address this evaluation and resources gap by introducing HeSum, a novel benchmark dataset specifically designed for Hebrew abstractive text summarization. HeSum consists of 10,000 article-summary pairs sourced from Hebrew news websites written by professionals. Linguistic analysis confirms HeSum's high abstractness and unique morphological challenges. We show that HeSum presents distinct difficulties even for state-of-the-art LLMs, establishing it as a valuable testbed for advancing generative language technology in Hebrew, and MRLs generative challenges in general.[1]

1 Introduction

Recent advances with large language models (LLMs, Brown et al., 2020; Chowdhery et al., 2023) demonstrate impressive capabilies, encompassing diverse tasks such as natural language (NL) understanding and reasoning, including *classification* tasks as commonsense reasoning (Bisk et al., 2020) and sentiment analysis (Liang et al., 2022), as well as *generative* tasks like summarization and dialogue systems (Thoppilan et al., 2022). However, these impressive demonstrations are primarily confined to the English language. Our understanding of how these models perform on low-resource languages is limited, as current testing primarily focuses on languages with abundant data (Ahuja et al., 2023; Lai et al., 2023). This concern is particularly relevant for morphologically rich languages (MRLs) such as Hebrew, which is known for their processing difficulty (Tsarfaty et al., 2019, 2020).

Despite advancements in natural language processing for Hebrew, which so far covered tasks as reading comprehension (Keren and Levy, 2021; Cohen et al., 2023), named entity recognition (Bareket and Tsarfaty, 2021), sentiment analysis (Chriqui and Yahav, 2022), and text-based geolocation (Paz-Argaman et al., 2023); a crucial gap persists in the ability to generate new, human-like text, as is required by *abstractive text generation*. Abstractive text-generation requires not only natural language understanding and reasoning over the input, but also the ability to create grammatically correct, and in particular *morpho-syntactically* correct and *morpho-semantically* coherent, fluent text that maintains consistent meanings. Notably, text-generation models are also prone to 'hallucinations' — generating factually incorrect content. These challenges are further amplified in Hebrew due to its morphological richness which leads to a complex realization of sentence structure and meaning.

In order to enable empirically quantified assessment of these aspects of text generation in MRLs, we present a novel benchmark dataset for **He**brew abstractive text **sum**marization (HeSum). HeSum consists of 10,000 pairs of articles and their corresponding summaries, all of which have been sourced from three different Hebrew news websites, all written by professional journalists. This curated collection offers several key advantages: (i) *High Abstractness* – extensive linguistic analysis validates HeSum's summaries as demonstrably more abstractive even when compared to English benchmarks. (ii) *Unique Hebrew Challenges* – linguistic analysis meticulously pinpoints the inherent complexities specific to Hebrew summarization, offering valuable insights into the nuanced characteristics that differentiate it from its English counterpart. (iii) *Thorough LLM Evaluation* – we conducted a comprehensive empirical analysis using

[1]The dataset, code, and fine-tuned models will be made publicly available upon publication https://github/anonymous.

Set	Size	Vocabulary size (over Articles)		Avg. Document Length		Avg. Coreference	Avg. Construct state	Article-Summary Semantic Similarity
		Lemmas	Tokens	Article	Summary	Article	Summary	BertScore
Train	8,000	47,903	269,168	1,427.4	33.2	98.8	2.4	0.76
Validation	1,000	23,134	104,383	1,410.0	33.8	87.9	2.5	0.76
Test	1,000	22,543	102,387	1,507.6	34.7	95.7	2.6	0.74

Table 1: Linguistic Analysis of the HeSum dataset.

state-of-the-art LLMs, demonstrating that HeSum presents unique challenges even for these sophisticated models. By combining high abstractiveness, nuanced morphological complexities, and a rigorous LLM evaluation, HeSum establishes itself as a valuable testbed for advancing the frontiers of abstractive text summarization in MRL settings.

2 The Challenge

Linguistic Challenges in Hebrew Morphologically rich languages (MRLs) pose distinct challenges for generative tasks, above and beyond Morphologically improverished ones as English.

In MRLs, each input token can be composed of multiple lexical and functional elements, each contributing to the overall structure and semantic meanings of the generated text. For instance, the Hebrew word 'וכשמביתנו' is composed of seven morphemes: 'ו' ('and'), 'כש' ('when'), 'ם' ('from'), 'ה' ('the'), 'בית' ('house'), 'של' ('of'), and 'אנחנו' ('us'). This has ramifications for both the understanding of MRL texts, a process that necessitates morphological segmentation, and for generating MRL texts, requiring morphological fusion. At comprehension, Hebrew poses an additional challenge due to its inherent ambiguity, with many tokens admitting multiple valid segmentations, e.g., 'הקפה' could be interpreted as 'ה'+'קפה' ('the'+'coffee'); as 'הקפה' ('orbit'); or as 'הקף' + 'של' + 'היא' ('perimeter'+'of'+'her'). During generation, the emergence of unseen morphological compositions, where unfamiliar morphemes combine in familiar ways, poses an additional challenge (Hofmann et al., 2021; Gueta et al., 2023). These challenges, coupled with inherent linguistic features as morphological inflections, construct-state nouns (*smixut*), and more, create a multifaceted challenge for LLMs in processing and generating Hebrew texts.

The HeSum Task We aim to unlock the comprehension-and-generation challenge in MRL settings by first tackling the abstractive text summarization task (Moratanch and Chitrakala, 2016), here focusing on Modern Hebrew.

Given an input document in Hebrew, specifically a news article, our goal is to generate a short, clear, summary of the key information in the Hebrew language. In contrast to abstractive summarization, here novel morphosyntactic structures need to be generated to communicate the summary.

3 Dataset, statistics and Analysis

3.1 Data Collection

The HeSum dataset consists of article-and-summary pairs. The articles were collected from three Hebrew news websites: "Shakuf"[2], "HaMakom"[3], and "The Seventh Eye" [4]. These websites focus on independent journalism, providing articles on topics such as government accountability, corporate influence, and environmental issues. Each article on these websites is accompanied by an extended subheading that serves as a brief summary of the content. To ensure data quality, articles that were not in Hebrew or had short summaries (i.e., the extended subheading was less than 10 tokens) were excluded from the dataset.

3.2 Linguistic Analysis

We examined the linguistic, syntactic, and semantic properties of the HeSum dataset. For the extraction of syntactic and semantic features, we utilized DictaBert (Shmidman et al., 2023). Additionally, AlephBert (Seker et al., 2022), a Hebrew-based BERT model (Devlin et al., 2018), was employed to compute semantic similarity between articles and their corresponding summaries, leveraging the BertScore method (Zhang et al., 2019). Notably, semantic similarity was performed only on article-summary pairs within the model's 512-token limit.

Table 1 highlights the Hebrew language's multifaceted complexities as reflected in this task. The notable disparity in the vocabulary size between token and lemma counts underscores extensive morphological richness, necessitating models adept

[2]https://shakuf.co.il
[3]https://www.ha-makom.co.il
[4]https://www.the7eye.org.il

Dataset	novel n-grams				CMP	RED (n=1)	RED (n=2)
	n = 1	n = 2	n = 3	n = 4			
CNN/Daily Mail	13.20	52.77	72.2	81.40	90.90	13.73	1.10
XSum	35.76	83.45	95.50	98.49	90.90	5.83	0.16
HeSum	42	73.2	82	85.36	95.48	4.83	0.104

Table 2: HeSum's Intrinsic Evaluation compared to English Benchmarks (CNN/Daily Mail and XSum).

at handling linguistic diversity. The abundance of morphological anaphoric expressions (coreferences) and numerous Hebrew construct state constructions necessitate advanced models attuned to contextual relationships and Hebrew's unique morphological traits. Additionally, the document length in this corpus necessitates models equipped for long-form text processing. Moreover, the relatively high semantic similarity score indicates effective information distillation in the summaries.

3.3 Summerization Intrinsic Analysis

To assess the challenges of the HeSum summaries we used three established metrics: (i) *Abstactness (novel n-grams)* – the percentage of summary n-grams absent in the article (Narayan et al., 2018). (ii) *Compression Ratio (CMP)* – the word counts in summary (S) divided by the corresponding article (A): $CMP_w(S, A) = 1 - \frac{|S|}{|A|}$. Higher compression ratios indicate greater word-level reduction and, subsequently, potentially pose a more challenging summarization task (Bommasani and Cardie, 2020). (iii) *Redundancy (RED)* – measures repetitive n-grams within a summary (S) using the formula: $RED(S) = \frac{\sum_{i=1}^{m}(f_i - 1)}{\sum_{i=1}^{m} f_i}$ where m is the number of unique n-grams in the summary and f_i represents a frequency of specific n-gram within the summary.

Table 2 presents a quantitative analysis of HeSum's summarization characteristics, underscoring its challenges. HeSum demonstrates a high degree of abstractness, with approximately half of its unique vocabulary and over 73% of bigrams absent from the original articles. Furthermore, HeSum presents a significant compression challenge, as summaries average less than 5% of the input article length. Additionally, the analysis reveals minimal redundancy within the summaries, with less than 5% repeated n-grams. These findings underscore HeSum's efficacy in conveying the central ideas of the articles' information in a novel, distillate, and non-redundant manner. Comparative analysis with established abstractive summarization benchmarks, CNN/Daily Mail (Nallapati et al., 2016) and XSum (Narayan et al., 2018), confirms HeSum's high abstractness, compression ratio, and low redundancy, even when compared to these datasets.

4 Experiments

4.1 Experimental setup

Models To demonstrate the complexity of this task, we conducted an evaluation of two LLMs in a zero-shot setting: the GPT-4 model with 32K context window (version 0613), and GPT-3.5-turbo with 16K context (version 0613). To find the most effective prompt format, we tested on the HeSum validation set various prompting strategies, including translating parts of the prompt to English. Ultimately, we adopted the English-translated approach (Brown et al., 2020), where both the instruction and input were translated. The output summaries are strictly in Hebrew. Additionally, due to the absence of available generative models for Hebrew, we fine-tuned the multilingual mT5 sequence-to-sequence model (xue, 2020) on the HeSum training set. Appendix B includes GPT models' prompting strategies experiments, and mT5 training details.

Automatic metrics To evaluate the generated summaries with respect to the original texts, we used two automatic metrics: Rouge and BertScore. Rouge (Lin, 2004) is a widely-used metric in summarization that measures n-gram overlap between generated summaries and human-written references. We calculated Rouge-1 (unigrams), Rouge-2 (bigrams), and Rouge-L scores (longest common subsequence) to capture different levels of granularity. However, n-gram metrics like Rouge can struggle with capturing semantic similarity if paraphrases are used. To address this, we also employed BertScore (Zhang et al., 2019) with AlephBert (Seker et al., 2021) as its backbone. BertScore leverages pre-trained language models to provide a more semantically meaningful evaluation.

Model	ROUGE			BertScore	Human Evaluation	
	ROUGE1	ROUGE2	ROUGEL		Coherence	Completeness
GPT-4	10.3	2.64	10.5	0.773	4.00	3.86
GPT-3.5	11.5	2.3	9.6	0.77	4.12	3.62
mT5 (fine-tuned)	12.8	4.26	11.6	0.5756	3.48	2.87

Table 3: Models' performance on the HeSum test-set.

Phenomenon	GPT-4	GPT-3.5	mT5	Example error in Hebrew	Example error translated into English	Explanation
Repetition	0	0	5	האם הוא יכול להיות אלים? אם הוא יכול להיות אלים?...	Can he be violent? If he can be violent?	Duplication with subtle alterations.
Token-merge	0	0	2	...ראש הממשלה אמרעוד פעם...	...the Prime Minister saidagain...	'saidagain' should be two words – 'said' and 'again'.
Hallucination	3	2	0	...עירב את נח...	...involved Noah...	Noah is not a person mentioned in the article.
Culture transfer	1	1	0	...למנהיגת הקמפיין הנבחרת, ננסי ברנדס...	...to the campaign leader-elect, Nancy Brands...	The article refers to Nancy as a 'he', but the summary uses feminine inflection (leader), probably due to Nancy being a common female name in English.
Incorect gender	4	7	0	...חושפות בחקירתם...	...reveal in their investigation...	Gender inflection mismatch: 'reveal' (fem.) clashes with 'their' (masc.).
Incorrect definite (e.g., construct state)	2	2	3	...המשרד המשפטים פירסם...	The Ministry of the Justice published...	Definite articles on both words in 'The Ministry of the Justice' violate Hebrew construct state rules.

Table 4: Error analysis comparing generated summaries from GPT-4, GPT-3.5, and mT5 based on 20 inputs.

Human Evaluation To validate the quality of model-generated summaries for the HeSum task, seven independent expert annotators evaluated a total of 186 summaries (62 per model) based on the same set of 62 reference articles. Annotators evaluated each summary using a 1-5 Likert scale (Likert, 1932) based on two key criteria: *coherence*, which assessed the summaries' grammaticality and readability, and *completeness*, which measured the degree to which they capture the main ideas of the articles. To measure the consistency of the annotators' scores, we calculated Krippendorff's α (Krippendorff, 2018) for an interval scale, and received an α score of 0.778 which indicates a good inter-annotator agreement rate.

4.2 Results

Quantitative Analysis Table 3 presents the quantitative evaluation results. On surafce similarity metrics (ROUGE), mT5 surpassed the GPT-based models. Notably, the ROUGE scores for GPT-based models on Hebrew lag behind other MRLs (Lai et al., 2023; Ahuja et al., 2023) on the abstractive summarization task, underscoring the NLP challenge in Hebrew. Interestingly, the GPT-based models exhibited an inverse trend, outperforming mT5 on the semantic similarity measure (BertScore). Furthermore, high-quality human evaluation, revealed that despite not being trained on the specific data, the GPT-based models achieve higher scores in both coherence and completeness.

Qualitative Analysis Following the identification of key error categories, we conducted a comparative analysis by randomly selecting 20 summaries generated by each of the three models for the same set of 20 articles. For each model, we then quantified the occurrences of each identified phenomenon within the sampled summaries. The results in Table 4 reveal disparities between the GPT-based models and the fine-tuned mT5 on various linguistic phenomena. The finetuned mT5 exhibits pronounced disruptions like repetition (20%) and token merge (10%), which weren't observed in the GPT-based results. However, the GPT-based models demonstrate errors in morphological phenomena specific to Hebrew, such as incorrect gender and wrong definiteness marking on *smixut*, indicating that the morphological richness of the language remains a challenge for these LLMs. Additionally, known phenomena of GPT-based models such as "hallucinations" (Cui et al., 2023; Guerreiro et al., 2023) are also seen in our analysis.

5 Conclusion

This research seeks to fill a critical gap in the field of assessing generative LLMs for MRLs by presenting HeSum, a new dataset that includes 10K article-summary pairs sourced from professional journalists on Hebrew news websites. HeSum offers three key advantages: high level of abstractness in summarization, distinct challenges specific to the Hebrew language, and a comprehensive empirical assessment of LLMs using this dataset. By integrating these aspects, HeSum establishes itself as a valuable resource for researchers striving to push the boundaries of generative tasks, and specifically abstractive text summarization in Hebrew.

Limitations

Although we aspired to evaluate HeSum on a broad range of large language models (LLMs), our current analysis is limited to only two generative models. This might overlook newer models offering potentially superior performance. Additionally, resource constraints prevented us from investigating the behavior of these models in few-shot settings. Having acknowledged that, the timeliness of this resource is uncompromized, as it can be used with contemporary and future models alike, to track advances on this challenge. Furthermore, time and cost constraints restricted the human evaluation to a comparatively small sample size, with only 62 summaries assessed out of the 1,000 in the test set. Last, HeSum predominantly comprises articles from news websites, which may bias models' success in this task towards news-style writing, and may not fully capture the linguistic diversity across different genres and domains. The reason for selecting these domains specifically stems from our ability to obtain a permissive license for the resource, allowing open and free access by the community.

Ethics

Following the generous permission of 'Shakuf', 'HaMakom', and 'The Seventh Eye' — organizations committed to independent journalism, media scrutiny, and transparency in Israel — we were granted the valuable opportunity not only to access and analyze their published articles but also to publish the data for broader research use. This unique collaboration fosters open access and empowers other researchers to build upon the data extracted from their articles and our findings within Hebrew abstraction summarization, expanding knowledge in this important field. Also, we are guaranteed not to have offensive language or hate speech in our data. It should be borne in mind, however, that the opinions or biases reflected in these data may differ from other sources of information (news websites, social media, non-Hebrew news reports, and the like). So, the deployment of technology trained on this resource should be done with care.

References

2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, et al. 2023. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*.

Google Translate API. 2023. Google translate api v2 documentation.

Dan Bareket and Reut Tsarfaty. 2021. Neural modeling for named entities and morphology (nemo^2). *Transactions of the Association for Computational Linguistics*, 9:909–928.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Rishi Bommasani and Claire Cardie. 2020. Intrinsic evaluation of summarization datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8075–8096.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Avihay Chriqui and Inbal Yahav. 2022. Hebert and hebemo: A hebrew bert model and a tool for polarity analysis and emotion recognition. *INFORMS Journal on Data Science*, 1(1):81–95.

Amir Cohen, Hilla Merhav-Fine, Yoav Goldberg, and Reut Tsarfaty. 2023. Heq: a large and diverse hebrew reading comprehension benchmark. pages 13693–13705.

Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. 2023. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Nuno M Guerreiro, Duarte M Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André FT Martins. 2023. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Eylon Gueta, Omer Goldman, and Reut Tsarfaty. 2023. Explicit morphological knowledge improves pre-training of language models for hebrew. *arXiv preprint arXiv:2311.00658*.

Valentin Hofmann, Janet B Pierrehumbert, and Hinrich Schütze. 2021. Superbizarre is not superb: Derivational morphology improves bert's interpretation of complex words. *arXiv preprint arXiv:2101.00403*.

Omri Keren and Omer Levy. 2021. Parashoot: A hebrew question answering dataset. *arXiv preprint arXiv:2109.11314*.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

N Moratanch and S Chitrakala. 2016. A survey on abstractive text summarization. In *2016 International Conference on Circuit, power and computing technologies (ICCPCT)*, pages 1–7. IEEE.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*.

Tzuf Paz-Argaman, Tal Bauman, Itai Mondshine, Itzhak Omer, Sagi Dalyot, and Reut Tsarfaty. 2023. Hegel: A novel dataset for geo-location from hebrew text. *arXiv preprint arXiv:2307.00509*.

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.

Amit Seker, Elron Bandel, Dan Bareket, Idan Brusilovsky, Refael Shaked Greenfeld, and Reut Tsarfaty. 2021. Alephbert: A hebrew large pre-trained language model to start-off your hebrew nlp application with. *arXiv preprint arXiv:2104.04052*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Shaltiel Shmidman, Avi Shmidman, and Moshe Koppel. 2023. Dictabert: A state-of-the-art bert suite for modern hebrew. *arXiv preprint arXiv:2308.16687*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Reut Tsarfaty, Dan Bareket, Stav Klein, and Amit Seker. 2020. From spmrl to nmrl: What did we learn (and unlearn) in a decade of parsing morphologically-rich languages (mrls)? *arXiv preprint arXiv:2005.01330*.

Reut Tsarfaty, Amit Seker, Shoval Sadde, and Stav Klein. 2019. What's wrong with hebrew nlp? and how to make it right. *arXiv preprint arXiv:1908.05453*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A The HeSum Dataset

Collection Protocol Since the websites we collected (Shakuf, HaMakom, and The Seventh Eye) lack archives or RSS feeds, we developed a crawler to systematically navigate through pages, beginning from the homepage and exploring various article links. Leveraging their shared HTML structure, we could efficiently scrape the sites. We excluded pages without textual content, such as multimedia pages or those not in Hebrew. Additionally, articles with summaries of less than 10 tokens were filtered out, as they often lack sufficient detail to be a summary. In addition, all the articles were cleaned from Unicode characters or unrelated content.

Coherence

1. Very Incoherent: The summary is extremely confusing and lacks any clear connection between sentences.
2. Incoherent: The summary is somewhat understandable.
3. Somewhat Coherent
4. Coherent
5. Very Coherent

Completeness

1. Very Incomplete: The summary lacks essential information and does not convey the main points effectively.
2. Incomplete: The summary provides some information but misses key details.
3. Somewhat Complete
4. Complete
5. Very Complete

Figure 1: Evaluation Criteria

Human Evaluation Details We collected annotations from seven volunteered participants aged 25 and above, all with at least one academic degree. The participants were instructed to rate two parameters – *coherence* and *completeness*, based on known criteria, as depicted in Figure 1.

B Models

Fine-tunning mT5 details For fine-tuning mT5, we utilized Google Colab's premium account, leveraging an open-source training code [5] for streamlined execution.[6] Training was conducted for three epochs on an A-100 GPU. We fine-tuned both mT5-small (300 million parameters) and mT5-base (580 million parameters) variants, with subsequent evaluation focused on mT5-base for its superior performance on the HeSum validation. Table 5 reports the mT5-base model's performance on various metrics on the HeSum validation set.

[5] https://github.com/imvladikon/hebrew_summarizer

[6] The fine-tuned model can be found at https://huggingface.co/hesum-anonymous/HeSum-mT5-base

Feature	Eval metrics
ROUGE1	25.77
ROUGE2	10.095
ROUGEL	19.88
Run Time	2.39
Loss	2.36
Samples	1000

Table 5: mT5-base performance on the validation set.

Model	prefix	input	output	ROUGE1	ROUGE2	ROUGEl
GPT-3.5	E	E	E	13.1	2.32	11
GPT-3.5	H	H	H	13	3	11.8
GPT-3.5	E	E	H	12.8	2.3	11
mT5	—	H	H	12.78	4.35	11.56
GPT-3.5	E	H	H	11.8	3	10.9
GPT-3.5	H	H	E	10.8	1.5	9.6
GPT-3.5	E	H	E	9.2	1.4	7
GPT-3.5	H	E	H	8	1	7
GPT-3.5	H	E	H	8	1	7

Table 6: Testing different configurations of language prompting to find the best configuration to evaluate GPT-3.5. 'H' denotes Hebrew and 'E' denotes English. 'prefix' is the instruction to the model, 'input' is the article itself, and the output is the desired summarization language.

You are a genius summarizer. Your task is to summarize the main points of the following text. Please follow these instructions step by step:

1. The summary should be concise, consisting of up to 3 sentences.
2. If there are several main topics, create a separate sentence for each topic.
3. The output should be in English.

Figure 2: The prompt we used for the GPT-based models

Prompting GPT-based models Here, we leverage the translate-English approach, suggested by (Shi et al., 2022) and (Ahuja et al., 2023), which translates instances from target languages into English before prompting. We decompose the prompt task into three parts: (i) the input article (ii) the

Dataset	novel n-grams				CMP	RED (n=1)	RED (n=2)
	n = 1	n = 2	n = 3	n = 4			
HeSum	42	73.2	82	85.36	95.48	4.83	0.104
GPT-4	47.24	80.35	91.37	95.92	91.89	8.14	0.68
GPT-3.5	45.69	80.18	91.73	96.35	93.46	7.53	0.83
mT5-finetuned	10.80	41.41	56.28	68.99	94.47	26.13	20.17

Table 7: Intrinsic Evaluation of Summarization. A Comparative Analysis of GPT-4, GPT-3.5, mT5 Models and the Hesum Dataset.

instruction (prefix), and (iii) the output. All three parts could be done in either Hebrew or English for the HeSum task. In our experiment, Google Translate API (2023, API, 2023) handled the translation of prompts (input and/or prefix) from Hebrew to English and the translated outputs back to Hebrew for analysis. Testing GPT3.5 on different configurations of language prompting in the HeSum validation set, we found that the best prompt-language configuration is English-English-English (Table 6). We then applied this prompting strategy to both GPT-3.5 and GPT-4 on the test set. The prompt we used depicted in Figure 2.

C Additional Models Performance Analysis

Table 7 presents the intrinsic evaluation results for the models, corresponding to the metrics introduced in Section 3.3. Notably, GPT-based models generate text with greater abstractness, as evidenced by their higher count of novel n-grams compared to the fine-tuned mT5. This finding aligns with mT5's tendency towards repetitive generation, which is further supported by its high RED score and by the qualitative analysis presented in Table 4.

D Implementation Details

For the intrinsic evaluation of the dataset, we created a Jupyter notebook which computes the different metrics. For computing the n-grams, we used the NLTK package,[7] and for loading and processing the data, we used NumPy[8] and Pandas.[9] For evaluation of the different models, we used the Rouge package [10] for ROUGE and Transformers[11] for BertScore.

[7] https://pypi.org/project/nltk/
[8] https://pypi.org/project/numpy/
[9] https://pypi.org/project/pandas/
[10] https://pypi.org/project/rouge/
[11] https://pypi.org/project/transformers/