INSS Insight

Does Israel Need a National Language Model?

Ariel Sobelman, Michael Genkin | No. 2047 | October 21, 2025

The world of artificial intelligence increasingly relies on large language models and chatbot applications. Through them, end users—from private individuals to government agencies—collect, generate, and consume most of the knowledge needed to manage their daily lives. Advanced nations are investing in building national language models that embody their culture, values, and national narratives. Israel has not yet developed such a model and must therefore rely on foreign systems and the narratives embedded within them. Just as the revival of the Hebrew language was an essential element in shaping Zionist and Israeli national identity, consideration should also be given to developing a Hebrew national language model that would be a central component of Israel's digital sovereignty.

A well-known joke in linguistics and the study of language claims that the difference between a dialect and a language is that a language is a dialect with an army. The meaning behind this jest is that while countless dialects exist in the world, only a much smaller number become official, recognized languages. A dialect is a means of communication serving a community—sometimes a large one—but one that has not crossed the elusive threshold to become part of a national identity tied to a political entity. One of many examples is the Basque language, which is widespread and forms part of the national identity of its speakers, yet lacks formal political standing—or, as linguists joke, since it has no army, it remains a dialect, which is a relatively lower category than an official language.

This insight is also relevant in the age of artificial intelligence (AI), whose foundation includes language models that power, among other things, "chatbot" systems such as ChatGPT and similar tools that have become part of everyday life in the past two years. AI and the era of large language models are merely another manifestation of the same principle, and in many ways, they lie at the core of national identity and digital sovereignty.

Recently, the Nagel Committee's report on accelerating AI development in Israel was published. In passing, the committee also addressed the question of a national language model—whether Israel needs one and whether it has the economic and technological capacity to develop it. On an abstract and philosophical level, this question deserves far greater prominence, since, in many respects, language is the foundation. However, first, it is necessary to explain what language models are and the technological, infrastructural, and economic challenges involved in developing them.

A large language model (LLM) is, in fact, a type of AI system that uses an enormous amount, between tens to hundreds of trillions of words worth of textual data to create a computerized representation of language capable of imitating speech and communicating with humans in a way that feels natural. During the "training" stage, that is, the conversion of textual data into a model, there is a preference for manually and organically created content gathered from diverse sources, including literature, journalism, and online publications, rather than machine-

generated text. This ensures more natural interactions. The system "infers" things in a way that appears rational, although its reasoning ability is statistical, based on billions of interactions, and attempts to provide the answers most likely to match the user's intent. AI systems, at least for now, are incapable of true rational thought or developing genuinely original ideas.[i] Despite these limitations, language models can conduct increasingly sophisticated human-like interactions, earning a certain degree of user trust.[ii] Commercial chatbot products often record and store conversations for later use as textual data to train new and more advanced models.

When considering the enormous amount of data that must be collected and processed to train large language models, it becomes easy to understand why the development of AI requires extensive computational and processing infrastructure, typically housed in massive data centers. These data centers require vast amounts of energy to operate and cool computers during the model-training and processing phases. Given the considerable cost of such infrastructure, it is hardly surprising that much of the committee's attention and most of the public discussion of its conclusions have focused on the budgetary requirements for accelerating and advancing this field in Israel.[iii] These demands amount to a sum exceed the current budget of Israel's national AI program.

A language model bases its responses, to a large extent, on the information it has been fed through human interaction. The choice of information and its content is, naturally, subjective. Simultaneously, the models also use objective information such as scientific data. For instance, a simple factual question such as "Does the sun rise in the east and set in the west?" is likely to yield an accurate answer regardless of the model queried. However, when it comes to subjective information, the model may assign disproportionate weight to the data on which it was trained and, in "good faith," offer an incorrect answer. This complexity becomes even more significant in chatbot applications such as ChatGPT or DeepSeek, because beyond the model itself is an additional layer of system instructions that guide the model on how to respond to user queries. When questions are more complex, especially those involving the analysis of value-based, cultural, national, or historical issues, these systems are exposed to various biases. Such biases are common in responses to contemporary or political topics, where the nature of the answers may be influenced by the lack of up-to-date information during training or by the developer's choice to omit certain data (or by the system instructions in the case of chatbot applications). As a result of these biases, it is clear that large language models (and chatbots even more so) reflect the worldview of their creators. Moreover, even if a model has been trained in multiple languages and can translate between them, without a deliberate effort to diversify its training data, it will generally reflect an Anglophone worldview because the most easily accessible content for model training is in English,[iv] and, by extension, this reflects the perspectives of English-speaking developers.[v]

Another potential source of bias and inaccuracy lies in the directives governing what content and subject areas are permitted or prohibited—restrictions typical of non-democratic regimes. For example, if one were to ask a chatbot developed by a Western company about the 1989 Tiananmen Square massacre in China, one would likely receive a range of answers referring explicitly to the massacre in the context of a civilian struggle against the Communist regime in the pursuit of democratic reform. Conversely, if one asked a chatbot developed by a Chinese company, such as DeepSeek, the response would necessarily conform to the official

state narrative at the expense of historical accuracy, or it would explicitly state that the chatbot cannot answer questions on that topic.[vi] A similar phenomenon, although less pronounced, can also occur when using the language models themselves and not only the chatbot applications built upon them.[vii] In any case, it is clear that language models are effectively part of a national system and serve as conduits for content that reflects their developers' identities. Just as formal spoken language is a component of identity and an element of nationhood and sovereignty, a large language model is also directly tied to the digital dimension of that same nationhood and sovereignty.

The central question of what kind of language model should be used and whether Israel should aspire to develop its own national language model was given only marginal attention in the Nagel Committee's report. At present, nine of the ten most advanced (frontier) language models have been developed by Western, primarily American, companies, such as OpenAI's GPT-4.5, Meta's Llama, xAI's Grok, Google's Gemini, and Anthropic's Claude, while only one frontier model was developed by a Chinese company. The cost of developing a frontier model is enormous, reaching tens of millions of dollars for the training phase alone. For example, the training cost of GPT-4.0 is estimated at around $50 million, excluding the capital expenses of constructing data centers, infrastructure, and the R&D needed to design the model—costs that can push total development expenses into the billions of dollars. Current estimates suggest a total cost of $850 million to over $1 billion for Meta's Llama 3.1 405B and $6–$9 billion for xAI's Grok 4.[viii]

Another potential limitation of creating a national language model is the need for a vast quantity of high-quality source text data. It is by no means certain that the Hebrew language possesses the tens of trillions of written words required to train a large language model from the ground up. One possible solution to this challenge would be to use an existing foundation model and adapt it to the Hebrew language through a process known as fine-tuning, which introduces the curators' own biases, although this can sometimes correct for biases in the original model. A successful example of this approach is DictaLM 2.0, a model adapted to Hebrew by training on approximately 50 billion Hebrew words drawn from textual sources. It achieves results that are superior to those of much larger models.[ix] Being an open model (meaning it can be deployed on computing infrastructure owned by the user rather than relying on the servers of the original developer), DictaLM allows its users to ensure that sensitive data processed by the model remain within Israel's borders. Among the frontier models, only one offers this capability, but its immense size makes it more computationally demanding. This is no magic solution; the model's size, and therefore its performance, are constrained by the computing resources available to its developers. With greater access to such resources, it would be possible to adapt a larger model—potentially even a frontier model such as Llama 3.1 405B—and produce a national model with even better results.

The importance of a national model can be illustrated by the Chat ha-Mishpat ("Judicial Chat") initiative recently announced by Israel's judicial authority. Under this initiative, the judiciary has developed and plans to deploy a chatbot system for approximately 900 judges, registrars, and their assistants, designed to streamline the management of court cases.[x] According to media reports, the Chat ha-Mishpat system is based on Google's Gemini model, which was purchased and specially adapted for the project. It operates as a closed system for each case (as it appears in the Legal Net [Net ha-Mishpat] judicial management system), reducing the

risks of data leakage or fabricated information, and reportedly achieves approximately 80% accuracy. The system is expected to significantly reduce the time required to prepare for hearings, aligning with the Nagel Committee's vision of leveraging AI to improve public sector efficiency. However, using Gemini for this initiative means that Israeli citizens' data may need to leave the country's borders to improve the efficiency of the judicial system. Moreover, processing costs are at least 25% higher, and the system's overall effectiveness is somewhat lower than that which could potentially be achieved with a dedicated national language model.[xi]

Therefore, a national language model can be profoundly significant for safeguarding the data sovereignty of Israeli citizens, improving the quality of responses provided by AI systems, preserving national culture and values, and ultimately contributing to Israel's national security. Nevertheless, it is impossible to separate the questions of such a model's necessity and value from the issue of its cost, although several approaches to its development could substantially affect that cost. On the face of it, one can state clearly that a national language model would support Israel's national narratives far more effectively than foreign ones. Linguistic nuances that form part of Israel's national identity would likewise be better reflected in domestic models. In the spirit of the linguists' half-joking metaphor, it is evident that national identity in the age of AI is directly linked to the existence of a national digital language—or a national language model. Therefore, despite the expense, developing such a model should be considered seriously.

In conclusion, a shared language has always been a key marker of human communities. This holds true for natural languages and equally for emerging digital languages. In the era of AI, Israel must not outsource the maintenance of its language—with all its cultural, moral, and identity-based dimensions—to foreign language models. As in other strategic domains, a degree of self-reliance is essential to avoid dependence on external systems. In recent years, much has been said about Israel's digital sovereignty. The authors propose that Israel's digital sovereignty fundamentally depends on the existence of a national large language model.

---

[i] Parshin Shojaee et al., "The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity," arXiv:2506.06941, arXiv, July 18, 2025, https://doi.org/10.48550/arXiv.2506.06941.

[ii] Joy Buchanan and William Hickman, "Do People Trust Humans More than ChatGPT?," *Journal of Behavioral and Experimental Economics* 112 (October 2024): 102239, https://doi.org/10.1016/j.socec.2024.102239.

[iii] Yaakov Nagel et al., *Report of the National Committee for Accelerating the Field of Artificial Intelligence – August 2025* (The National Committee for Accelerating the Field of Artificial Intelligence, 2025), 62–64 [Hebrew], https://www.gov.il/he/pages/event-ai050825; Shlomo Teitelbaum and Adrian Filut, "Nagel Committee: Israel's AI Situation Is Troubling, Recommends Massive Investments and Tax Incentives for Workers in the Field," *Calcalist*, August 6, 2025 [Hebrew], https://www.calcalist.co.il/calcalistech/article/skux4lgoex.

[iv] Pablo Villalobos et al., "Will We Run out of Data? Limits of LLM Scaling Based on Human-Generated Data," arXiv:2211.04325, arXiv, June 4, 2024, https://doi.org/10.48550/arXiv.2211.04325.

[v] Lisa Schut et al., "Do Multilingual LLMs Think In English?," arXiv:2502.15603, arXiv, February 21, 2025, https://doi.org/10.48550/arXiv.2502.15603.

[vi] Zeyi Yang, "Here's How DeepSeek Censorship Actually Works—and How to Get Around It," *Wired*, January 31, 2025, https://www.wired.com/story/deepseek-censorship.

[vii] Ali Naseh et al., "R1dacted: Investigating Local Censorship in DeepSeek's R1 Language Model," arXiv:2505.12625, arXiv, May 19, 2025, https://doi.org/10.48550/arXiv.2505.12625.

[viii] James Sanders, "What Did It Take to Train Grok 4?," Epoch AI, September 12, 2025, https://epoch.ai/data-insights/grok-4-training-resources; Louie Peters [@_LouiePeters], "LLama 3.1 405B likely cost ~$60m after training for ~100 days on a 16k H100 datacentre (maybe ~$850m capex). This is a huge free gift to open source AI & anyone building with AI! It is not yet however a next generation LLM (eg GPT-5, Claude Opus 3.5, Gemini Ultra 1.5/2.0, Grok-3," X, July 25, 2024, https://x.com/_LouiePeters/status/1816443587053092917.

[ix] Shaltiel Shmidman et al., Hebrew LLM Leaderboard, Python, DICTA, released March 12, 2025, https://huggingface.co/spaces/hebrew-llm-leaderboard/leaderboard; Shaltiel Shmidman et al., "Adapting LLMs to Hebrew:

Unveiling DictaLM 2.0 with Enhanced Vocabulary and Instruction Capabilities," arXiv:2407.07080v1, arXiv, July 9, 2024, https://doi.org/10.48550/arXiv.2407.07080.

[x] Netael Bandel, "Judge, Defendant, and Chatbot: This Is What the Trial of the Future Will Look Like," *Ynet*, August 10, 2025 [Hebrew], https://www.ynet.co.il/digital/technews/article/hkcucns00le; Itamar Levin, "900 Judges Will Soon Receive a New Tool to Assist Their Work, and the Reactions Are Enthusiastic," *Globes*, August 24, 2025 [Hebrew], https://www.globes.co.il/news/article.aspx?did=1001519639

[xi] "Israel Data Boundary," Google Cloud, https://cloud.google.com/assured-workloads/docs/control-packages/israel-data-boundary; "Llama 3.3 70B - Intelligence, Performance & Price Analysis," Artificial Analysis, https://artificialanalysis.ai/models/llama-3-3-instruct-70b.