Shahar Adar (207031311), Mattan Hoory (319003711)
Faculty of Data and Decision Sciences, The Technion
Project in NLP (00970215)
September 2025

MAPTIC: Evaluating Deterministic Mapping Capabilities of Large Language Models

We present MAPTIC, a framework for generating and evaluating datasets that test the deterministic mapping abilities of LLMs. Unlike most benchmarks that emphasize reasoning, MAPTIC targets exact input to output transformations (e.g., 1:"one"). Initial results show that even state-of-the-art LLMs make systematic mapping errors, pointing to the influence of tokenization artifacts, embedding limitations on domain specific sequences, sequence length, and task complexity. MAPTIC thus provides a scalable benchmark for assessing and improving the reliability of LLMs in such tasks.

Introduction. Life itself depends on mappings. Every three nucleotides, the basic building blocks of RNA and DNA, are translated into a single amino acid, building the proteins that power all living cells. This process is formalized in the *central dogma of molecular biology*[1]. Mapping is a universal idea: In mathematics, it appears as functions: deterministic rules pairing each input with a single output, forming the backbone of algorithms, databases, and programming languages. In real-world applications, *correctness* is non-negotiable: a single mapping mistake breaks an entire processing workflow. As LLMs move into production, their ability to perform such exact, reproducible transformations becomes almost as critical as their capacity for reasoning and serves as a test of their *compositional generalization*[2]

Dataset. We constructed a dataset of ten mapping tasks ("topics"), each with three difficulty levels (easy, medium, hard) that scale with input length. For each task-difficulty pair, 27 prompts were generated, yielding 810 examples in total, with code supporting scalable expansion. The chosen topics contained symbolic, linguistic, and biological mappings, providing a diverse method of evaluation:

Table 1. Samples per topic, prior to their embedding within prompt templates containing instructions.

Topic	RNA	Case	Country Code	Digits
Input	CCACCAAAG	CAT	AU DE	1 2
Expected Output	PPK	cat	Australia Germany	one two

1. Case Conversion: To evaluate token-level transformation abilities, we curated datasets containing randomly generated strings and natural text snippets (sampled from NLTK[3] corpora: Gutenberg, Reuters, WebText). Each example was converted from lowercase to uppercase (and vice versa). These tasks test whether models can handle in and out-of-distribution inputs, testing the influence of tokenization and pretraining on model performance.
2. RNA-to-Protein Mapping: Strings representing RNA sequences were translated into protein sequences by mapping RNA 3-mers to single-characters representing the 20 amino acids and stop signals (e.g. 'AUG':'M'). Input sequences were randomly generated with lengths proportional to difficulty, and outputs were computed deterministically via the codon table[4] widely available online.
3. Country Code Resolution: Country codes were sampled from ISO 3166-1 alpha-2[5] standard and mapped to corresponding country names. Easy examples used only single-word country names, while harder ones included multi-word entities, testing whether models can function as reliable lookup tables for standardized symbolic mappings.
4. Numbers to Digits: Integer digits (1-9) were converted to their corresponding English words.

Evaluation. We implemented a character- and element-level evaluator inspired by dynamic programming sequence-alignment algorithms[6,7] from computational biology. Using a Levenshtein[8]-style edit distance, outputs were aligned to pre-generated references, and each discrepancy was labeled as insertion, deletion, or substitution. This enabled fine-grained error analysis beyond binary accuracy.

Experimental Setup and MAPTIC Results. Due to compute constraints, the full dataset was evaluated on the Gemini 2.5 Flash model (no thinking), while a balanced subset of 20 examples per task-difficulty was sampled for evaluation with Gemini 2.5 Pro (limited to minimal thinking). Without additional tools.

Table 2. Mean per-element error rate (%) per topic across all prompts, where "element" is topic-dependent (e.g., character, 3-mer, or country name). A dash indicates no errors. Correct is the portion of perfect (all elements matched) answers. Bold marks the highest error in each row; underlines mark the better model per topic. Error type describes insertion/deletion/substitution proportions (I/D/S, %).

Task	Model	Easy	Medium	Hard	Avg.	Correct	Err. Type
To Lower (rand.)	flash	–	6.0245	**9.2188**	5.0811	61.7	14.6/44.9/40.5

The project was carried out under the supervision of Dr. Eyal Ben-David in the Faculty of Data and Decision Sciences, The Technion - Israel Institute of Technology. The code for reproducing the work is available at: https://github.com/mathoory/llm-mapping-dataset

To Lower (text)	pro	–	0.7143	**1.6807**	0.7983	80.9	57.1/14.3/28.6
	flash	–	**0.0374**	0.0069	0.0148	91.3	36.4/54.5/9.1
To Upper (rand.)	pro	–	–	**0.1302**	0.0434	95.6	0/75/25
	flash	1.4815	0.7231	**6.1935**	2.7994	74.1	14.4/65.6/20
To Upper (text)	flash	0.0176	0.0415	**0.0516**	0.0369	91.3	17.4/56.5/26.1
RNA to protein sequence	flash	44.3739	58.6436	**65.2747**	56.0974	1.2	12.1/21.1/66.7
	pro	30.7937	56.3679	**62.3727**	49.8448	10	8.3/30.6/61.1
Digits to number words	flash	–	–	**0.0281**	0.0094	87.6	99.5/0.4/0.1
	pro	–	–	**0.0004**	0.0001	85.7	33.3/0/66.7
Country codes to names	flash	0.0296	0.0815	**0.1074**	0.0728	42	12.4/0/87.6
	pro	–	0.0143	**0.0300**	0.0148	75	0/0/100

Table 3. Example errors for selected tasks from the easy category.

Task	Input	Expected Output	Model Output
Lower to upper (random)	idfmv	IDFMV	IDFVM
RNA to protein sequence	CGCUACCCCU GGAUA	RYPWI	RYPLD
Country codes to names	LU TN MW MO MS	Luxembourg; Tunisia; Malawi; Macao; Montserrat	Luxembourg; Tunisia; Malawi; Morocco; Mauritius

To investigate tokenization and embedding effects, we examined two of the error cases from Table 3. Although both the RNA and country-code sequences consist of five elements, they were segmented into nine and six tokens respectively. Such discrepancies indicate that meaningful units are not preserved as atomic tokens. Embedding analysis based on cosine similarity (Fig. 1) shows lack of specificity between 3-mers and amino acids, whereas country codes align more clearly with their respective names. This suggests that both tokenization artifacts and weaker embedding representations might contribute to the errors. Our analysis is constrained by the Gemini API, which exposes only token counts and embeddings rather than full token boundaries.

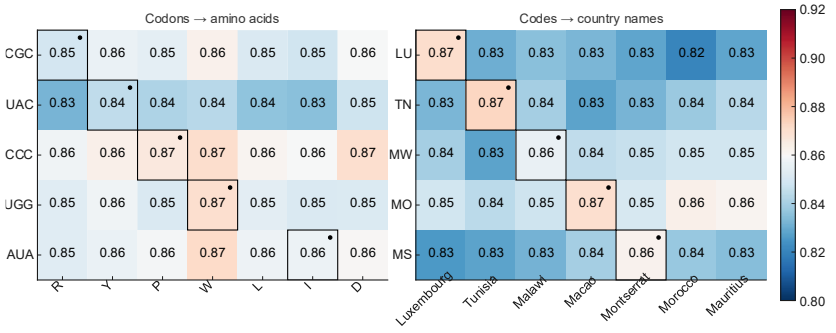

Figure 1. Cosine similarity heatmaps for the error cases presented in Table 3: (left) codons versus amino acids and (right) country codes versus country names. Bold entries mark the expected mappings.

Conclusions. Gemini 2.5 models exhibit errors across nearly all deterministic mapping tasks. Per-element error rates might understate the impact of a single mistake if it invalidates an entire prompt under stricter evaluation. The largest errors occurred in RNA-to-protein translation (>49%) and increased with sequence length, showing that domain-specific and compositional mappings remain particularly challenging. Task difficulty also scaled with the number of distinct mapping constraints: full alphabets were harder than digits. Error patterns suggest that tokenization granularity and embedding quality play a key role: higher error rates on random strings compared with natural text, and the asymmetry between lowercase-to-uppercase and its reverse, point to weaker representations for some token types. Conversely, near-perfect results on ISO country code mappings suggest that frequent, semantically rich tokens benefit from stronger embeddings. Error profiles vary by task and by model: Flash shows more deletions, while Pro produces more substitutions, indicating better sequence length preservation but imperfect token selection. Overall, Pro outperformed Flash, particularly on harder tasks, but neither achieved consistent production-ready responses. These findings highlight the importance of MAPTIC and motivate further work. The framework allows extending the dataset by scaling its size, refining the task set with harder mappings, and evaluating additional models, while future work may also include a systematic analysis of prompts, tokenization, and embedding effects.

References

1. CRICK, F. H. On protein synthesis. *Symp Soc Exp Biol* **12**, 138–63 (1958).

2. Chomsky, Noam. *Aspects of the Theory of Syntax*. (M.I.T. Press, 1965).

3. Bird, Steven., Klein, Ewan. & Loper, Edward. *Natural Language Processing with Python*. (O'Reilly, 2009).

4. Andrzej (Anjay) Elzanowski, J. O. The Genetic Codes. https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?chapter=tgencodes#SG1 (2024).

5. International Organization for Standardization (ISO). ISO 3166 Country Codes. https://www.iso.org/iso-3166-country-codes.html.

6. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**, 443–453 (1970).

7. Navarro, G. A guided tour to approximate string matching. *ACM Comput Surv* **33**, 31–88 (2001).

8. Levenshtein, V. I. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady* (1966).