קאלעם

October 6–9, 2026
San Francisco, CA

MameLoshnLM: Yiddish Language Model and Evaluation Benchmark

Uri Katz[1,4] Omer Goldman[2] Tomasz Limisiewicz[3] Reut Tsarfaty[1] Noah A. Smith[3,4]

[1]Bar-Ilan University [2]University of Cambridge [3]University of Washington [4]Allen Institute for AI

Why Yiddish? (and what you can learn from it)

A fusion language: Germanic grammar, Hebrew script, three lexical layers

דִּי בָּאָבע דערציילט אַ שיינע מעשׂה

Di bobe dertseylt a sheyne mayse

"Grandma tells a nice story"

Germanic Slavic Hebrew/Aramaic (Loshn-Koydesh)

Rich textual tradition
Centuries of literature, press and religious writing, most of it never digitized.

Thin, noisy web presence
Poorly represented in the web crawls used to train modern LMs, and easily confused with Hebrew, which shares its script.

Many languages share this profile: heritage-rich, web-poor, script shared with a bigger neighbor. Yiddish lets us measure what that does to an LM.

The first comprehensive LLM effort for Yiddish



Oytser Corpus
915M words of high-quality Yiddish
~6× mC4 ~58× OSCAR



MameLoshnLM
Continued pretraining of Llama 3.1 8B



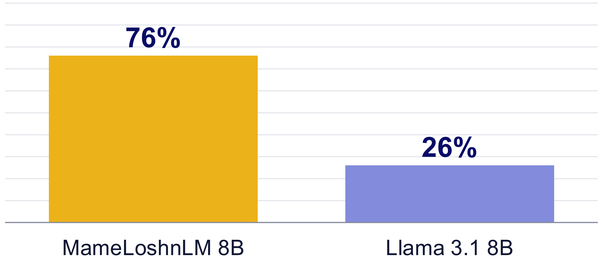
Kashes
14 Multi-task **evaluation benchmark**

Multilingual LMs can look fluent while missing what makes a language unique.

General-purpose models produce understandable Yiddish, but systematically underuse key linguistic phenomena.

Underproduction of a language-specific lexical stratum

Use of Loshn-Koydesh (Hebrew-origin) words, as % of the human rate.



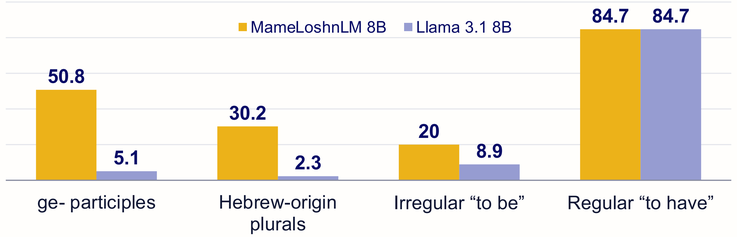
What does underproduction look like?

Meaning (English)	LK form (used by MameLoshnLM)	Llama alternative (e.g. Germanic)
Family	Mishpokhe	Familye
War	Milkhome	Krig
Face	Ponim	Gezikht

Meaning is often preserved, but a characteristic lexical layer of Yiddish is underused, and lost.

Morphological competence: large gaps on language-defining patterns

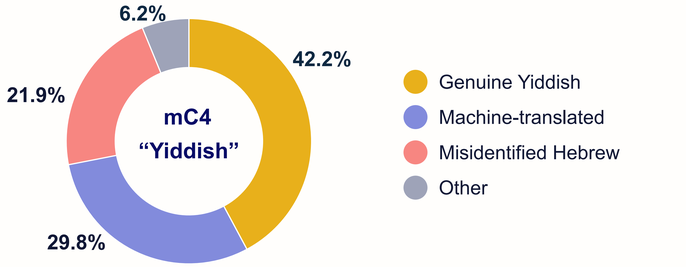
Exact match accuracy (%) on targeted morphological phenomena.



Some phenomena are learned, while language-defining morphology is largely missing from general-purpose LMs.

Web-scale "Yiddish" is not very Yiddish

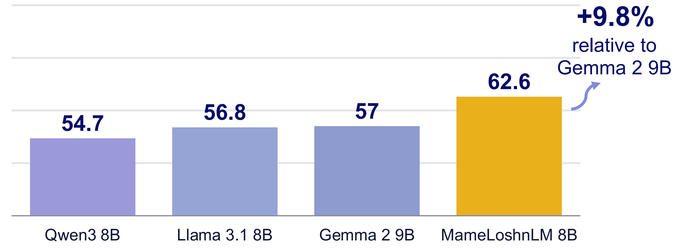
Our audit of mC4 shows that a large fraction of the data is noisy, machine-translated, or not Yiddish at all.



Noisy multilingual data may teach models a plausible approximation of a language rather than the language itself.

MameLoshnLM outperforms same size models across a broad benchmark

Average performance across 14 tasks.



On English → Yiddish translation, MameLoshnLM improves COMET by +13.6 over the next best baseline.

github.com/katzurik/MameLoshnLM

huggingface.co/Yiddish-NLP