

Small Language Models in Practice

Axel Menning
KIDA KON 2026

SLMs master language without being Encyclopedias

Coding models do not need Goethe's Faust.

Task-Bound Engines

- Performant on narrow, deterministic tasks.
- Acting as reasoning engines
- Tool-calls and external knowledge enhance capabilities
- Not covered in this talk:
 - Efficient in agentic systems
 - Fine-tunability (not covered in this talk)

Why explore for KIDA?

- ✓ CPU offload
- ✓ Zero network traffic
- Less administrative overhead
- Data sovereignty, Personal Information Conformity

SLMs might complement, not replace existing AI infrastructure at BfR

Current SLM Landscape

Open weights are the entire point: If it's not running locally, use a frontier model

	Footprint	Context Window
Gemma Family (Google)	4B • 12B	128K (E2B/E4B), 256K (12B)
Qwen Series (Alibaba)	2B • 4B • 9B	128K
Phi Family (Microsoft)	3.8B • 14B	128K
Llama Edge (Meta)	1B • 3B	128K
Bonsai Family (PrismML)	1.7B • 8B • 27B	262K
Ling Tiny (inclusionAI)	~1.3B Active (MoE)	262K
IBM Granite	3B • 8B	128K
LFM2 (Liquid AI)	1.2B • 2.6B • 7B	128K

How SLMs are created

Small models don't learn from scratch—they compress the wisdom of frontier architectures.

1. Precision & Weight Compression

Shrinking Weight Resolution & Network Size

Weights, Bit-Rate & Quantization FP16 (2B) $\rightarrow$ 4 Bit (0.5B) / 1.58 Bit

- Eg. 27B model in FP16 requires 54 GB VRAM just to load.
- Downscaling precision saves up to 75% VRAM.
- Ternary weights (-1, 0, 1) compress 27B models to run in <8 GB RAM.

Structural Pruning

- Physically trims layers and attention heads that only contribute minimally to accuracy.

Golden Rule: A larger model compressed down to 4-bit precision outperforms a smaller model running at full 16-bit precision within the same memory budget.

2. Data Quality & Distillation

Maximizing Information Density Per Parameter

Knowledge Distillation “*Textbooks are all you need*”

Textbooks: Curated high-signal reasoning data replaces raw, noisy web-crawl text.

Teacher $\rightarrow$ Student: Student model mimics the output probability distributions of a frontier LLM.

Domain Specialization **Focused Corpora**

- Sacrifices broad world trivia to achieve frontier reasoning in target tasks.

Efficiency Driver: 10 $\times$ higher token density enables 2B–7B models to match past 70B benchmarks.

Total Model Size (Bytes)

$$= N \text{ (Params)} \times B \text{ (Bits / Param)}$$

- **Knowledge Distillation & Pruning** reduce **N** by transferring capability into a physically smaller architecture (e.g., 70B $\rightarrow$ 3B / 1.5B).
- **Quantization** reduces **B** by lowering numeric precision (e.g., 16-bit float $\rightarrow$ 8-bit, 4-bit, or 2-bit).

How SLMs are deployed

Bringing Zero Infrastructure Inference to Ordinary Office Hardware and Edge Devices

Architectural shift: From centralized cloud GPUs to sovereign, distributed local endpoints

Consumer RAM • WebGPU • Edge NPUs

Quantized Silicon

1.5 – 8 GB

Extreme Scaling: Sub-4-bit model quantization

Standard RAM: Consumer workstation memory

Zero Discrete GPU: Democratized hardware access

In-Browser WebGPU

100% Offline

Zero-Install: Browser-native execution kernels

Data Sovereignty: Air-gapped local compliance

Zero Cloud Toll: Free from API network fees

Embedded Edge

Ultra-Low Watt

Hardware NPU: Mobile and tablet acceleration

Physical Edge: Offline vision and audio models

Lean Context: Recurrent KV-cache-free states

How SLMs are used

Replacing specialized traditional models where general world knowledge is not required.

TEXT & STRUCTURED TASKS

Language Processing

✓ Summarizing & Reformatting

✓ Translation

Structured output

✓ Semantic Search

✓ Relevance Ranking

✓ Named Entity Recognition

✓ Pointing out spelling & grammar mistakes

✓ Document Q&A

MULTIMODAL CAPABILITIES

Vision & Audio

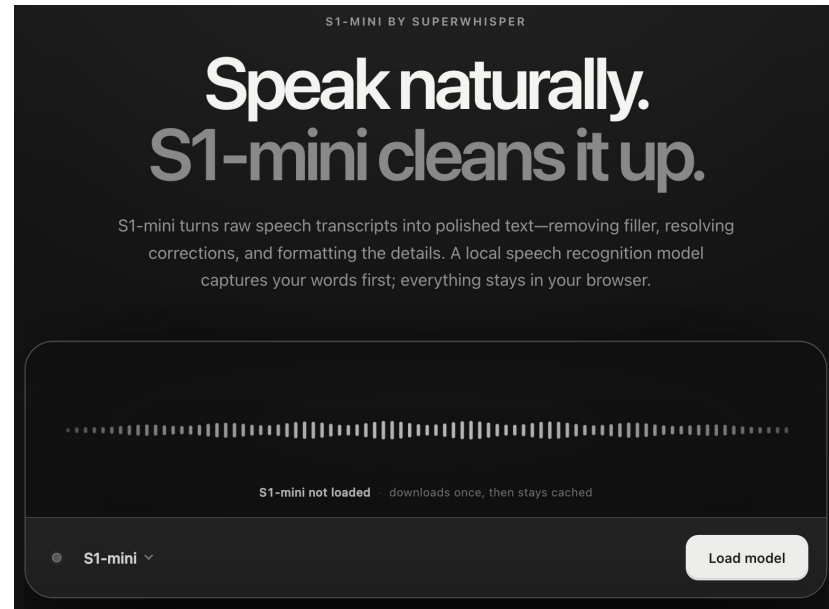
✓ Edge Vision and OCR: eg.: Qwen3.5-2B-MTP

✓ Speech-to-Text / Speech normalization

Examples 1: Structured Output and Text Normalizer



Osmosis-Structure-0.6B: Small Language Model for Structured Outputs



A 0.6B-parameter text normalizer for speech-to-text output:
<https://huggingface.co/spaces/webml-community/s1-mini-webgpu>

Examples 2: SLM for Search Without Heavy Embeddings

Two studies on making language models practical for search systems

STUDY 1 • LINKEDIN

Semantic Job Search

ARXIV:2510.22101

Goal: Rank search candidates in real-time under strict latency constraints (scoring 3M+ items/sec).

40%

Model Size Reduction (Pruning)

10x

Serving Throughput Gain

STUDY 2 • KANG ET AL.

Enterprise Search Labelers

ARXIV:2601.03211

Goal: Fine-tune small language models to create a fast, low-cost relevance labeler.

19x

Cost Efficiency Gain

17x

Annotation Throughput Increase

UseCase: ON-DEVICE KNOWLEDGE ARCHITECTURE

 **karpathy / llm-wiki.md**
Created 5 months ago • Report abuse

 Subscribe

 Star 5,000+

 Fork 5,000+

<> **Code**  Revisions 1  Stars 5,000+  Forks 5,000+ Embed  Download ZIP

llm-wiki

 `llm-wiki.md` Raw

LLM Wiki

A pattern for building personal knowledge bases using LLMs.

This is an idea file, it is designed to be copy pasted to your own LLM Agent (e.g. OpenAI Codex, Claude Code, OpenCode / Pi, or etc.). Its goal is to communicate the high level idea, but your agent will build out the specifics in collaboration with you.

<https://gist.github.com/karpathy/442a6bf555914893e9891c11519de94f>

LLM Wiki

Continuously compiled knowledge base architecture vs. real-time context retrieval (RAG)

The LLM Wiki Paradigm

A continuously compiled, human-readable repository.

Storage: Persistent Markdown repository

Deterministic Links: Explicit `[[wikilinks]]` preserve relational truth.

Synthesized Growth: Ingestion updates existing concepts automatically.

Knowledge compounded over time.

Ingest files

Ask Question

LLM_Wiki Engine

STRUCTURED REPOSITORY COMPONENTS

.md files

Page Metadata

Index

Summary

Schema

Named Entity

Log

Relationships / `[[wikilinks]]`

<https://github.com/axmng/slm-wiki>

The Local SLM Wiki

Thesis: The underlying data philosophy of an LLM-Wiki solves fundamental weaknesses of small models.

Core Architectural Benefits

Solving Small Language Model Constraints

1. Structured External Artifacts

Substitutes parameter capacity with persistent structured data.

2. Explicit Hierarchies

Replaces complex vector search with deterministic structures.

3. Context Optimization

Tames context window & eliminates "KV-Cache Tax".

4. Sovereign & Zero Cost

Complete local data sovereignty with zero marginal inference cost.

Algorithmic SLM Workarounds

Replacing fragile LLM tasks with deterministic code

Structured JSON Extraction

Used instead of unreliable in-place text rewrites.

Algorithmic Lemmatization

Ensures canonical entity names through deterministic rules.

Regex "Wikilink Weaving"

Automatic link creation requiring zero model bookkeeping.

Atomic Serialization

Guarantees state consistency across **INDEX.md** & **LOG.md**

System Architecture Flow

1. Map-Reduce Ingestion

STEP 1: INGESTION

Extraction & Summarizing

Distills documents into concise markdown notes without hallucinating facts.



STEP 2: NAVIGATION

Index-Based Routing

Scans master topic index to pinpoint relevant notes for questions.



STEP 3: EXECUTION

Grounded Synthesis

Answers queries using only supplied notes without requiring world knowledge.

2. Hierarchical Two-Pass Retrieval

STEP 01: MASTER SCAN

Index Routing

Scans compact **INDEX.md** summary table instead of parsing the entire vault.



STEP 02: PRUNING

Target Selection

SLM outputs top 2-3 specific filenames matching user question intent precisely.



STEP 03: SYNTHESIS

Grounded Answer

Loads selected files into 2k token context to produce verified, grounded answer.

Zero Context Drift

Axel.menning@bfr.bund.de

German Federal Institute for Risk Assessment
bfr.bund.de/en

CC-BY-ND 4.0

BfR | Identifying Risks –
Protecting Health

Consumer health protection to go
BfR2GO – the BfR Science Magazine

bfr.bund.de/en/science_magazine_bfr2go.html

Follow us

-  @bfrde | @bfren | @Bf3R_centre
-  @bfrde
-  youtube.com/@bfr_bund
-  social.bund.de/@bfr
-  linkedin.com/company/bundesinstitut-f-r-risikobewertung
-  soundcloud.com/risikobewertung

Backup

Ingest and Map-Reduce

 SLM Wiki

Gemma 4 (WebGPU)

Query & Synthesize

Ingest & Map-Reduce

Vault Explorer (2)

Gemma 4 (Google WebGPU) (~1.5...

IndexedDB (Persistent)

Wiki: Default +

Ingest & Map-Reduce Pipeline

Archive raw texts, PDFs, or DOCX documents into `raw/`, automatically extract topics with local SLM, update Markdown notes in `pages/`, and maintain `INDEX.md`.

Try sample: Transformers Neuromorphic

DESTINATION WIKI

Currently Active: Default

Ingest into existing wiki

Create brand new wiki for this doc

Name for the new wiki:

Testi

SOURCE TITLE (OPTIONAL)

Roschanski, VIM-1 Stall N2, 2019

DOCUMENT / ARTICLE CONTENT

Upload .txt, .md, .pdf, or .docx

Loaded PDF : Roschanski, VIM-1 Stall N2, 2019 (12 pages) 40200 characters

Detection of VIM-1-Producing Salmonella enterica Serovars Infantis and Goldcoast at a Breeding Pig Farm in Germany in 2017 and Their Molecular Relationship to Former VIM-1-Producing German Livestock Production Nicole Roschanski, a * Sead Hadziabdic, b Maria Borowiak, b Burkhard Malorny, Annemarie Kaesbohrer, b,c Sebastian Guenther, a * Istvan Szabo, a Institute for Animal Hygiene and Environmental Health, Freie Universitaet Berlin, Berlin, Germany

40200 characters (~10050 tokens)

Processing Ingestion Pipeline...

Synthesizing Markdown note (1/4): [[VIM-1]]...

56%

DISCOVERED CONCEPT TOPICS (4):

[[VIM-1]]
WGS data of these isolates were compared with those for the previously isolated VIM-1-producing nella Infantis isolates from pigs...

[[Serovars Infantis]]
Serovars Infantis represents a primary subject and functional mechanism analyzed in the source document.

[[Salmonella Enterica]]

[[ACC-1]]

Vault Explorer and Editor

The screenshot displays the SLM Wiki Vault Explorer and Editor interface. The top navigation bar includes the SLM Wiki logo, a dropdown menu for the Wiki (set to Default), and buttons for Query & Synthesize, Ingest & Map-Reduce, and Vault Explorer (10). The right side of the top bar shows the current model (Gemma 4 (Google WebGPU) (~1.5...)) and buttons for IndexedDB (Persistent) and Mount Local Folder.

The main interface is divided into two panels. The left panel, titled 'Filter vault notes...', contains a sidebar with a file tree. The tree is organized into sections: ROOT VAULT FILES (INDEX.md, LOG.md), PAGES/ (10) (ACC-1, CPE, Freie Universit, LB, MEM, Neural Networks, Quantum Computing), and RAW/ (3) (Pauly,method comparison, 2021, Roschanski, VIM-1 Stall N2, 2019, Initial Knowledge Seed). The 'Salmonella Enterica' entry is highlighted in blue.

The right panel displays the content of the selected 'Salmonella Enterica.md' file. It features a title bar with the file name and an 'Edit Markdown' button. Below the title bar, there is a section for 'Links (4):' with four links: [[VIM-1]], [[Serovars Infantis]], [[ACC-1]], and [[INDEX]]. The main content area contains the following text:

Salmonella Enterica

Executive Summary: Salmonella Enterica is a genus of Gram-negative bacteria responsible for causing various diarrheal and systemic infections through ingestion of contaminated food or water.

Core Mechanisms & Insights

Key Characteristics:

- Pathogenicity:** Causes invasive or non-invasive infections depending on the strain and host, often through the bacterium adhering to the intestinal mucosa and producing toxins or invading host cells.
- Transmission:** Primarily spread via the fecal-oral route, often involving contaminated food preparation, water sources, or cross-contamination in food handling environments.
- Virulence Factors:** Possesses various virulence factors, including hemolysins and exotoxins, which contribute to tissue damage and immune evasion.
- Survival:** Exhibits robust survival capabilities in the gastrointestinal tract and various environmental niches, facilitating colonization and subsequent infection.
- Implications:**
- Clinical Context:** Leads to gastroenteritis, characterized by symptoms such as fever, abdominal cramps, and diarrhea, which can cause dehydration and secondary complications.
- Public Health:** Highlights the critical need for stringent food safety regulations, proper hygiene practices in food processing, and rigorous water quality control to prevent outbreaks.
- Diagnostics:** Requires specific microbiological techniques (e.g., culture isolation, PCR) for accurate identification and strain typing in clinical and food safety surveillance.
- Food Safety:** Emphasizes the importance of proper food temperature control, cooking procedures, and sanitation protocols to ensure pathogen inactivation during food preparation.

query

 SLM Wiki

Gemma 4 (WebGPU)

Wiki: Default

+

Query & Synthesize

Ingest & Map-Reduce

Vault Explorer (10)

Gemma 4 (Google WebGPU) (~1.5...)

IndexedDB (Persistent)

Mount Local Folder

 Hierarchical SLM Query & Synthesis

Pass 1 scans **INDEX.md** to select relevant files, then Pass 2 synthesizes grounded answers strictly from those notes.

Engine: WebGPU: Gemma 4 (Google)

what kind of samples do I need for isolation of CPE strains?

Ask ↵

Suggestions:

What are the core principles of quantum computin...

How do neural networks connect to quantum algor...

Explain self-attention and transformer bottlenecks.

What is neuromorphic computing and spiking neur...

 PASS 1: DIRECT INDEX SCAN & FILE SELECTION

"The query asks what samples are needed for the isolation of CPE strains. CPE.md directly addresses isolation limitations for CPE detection, and MEM.md describes the selective enrichment process involving relevant antibiotics (CTX and meropenem) used for isolating coli, which is relevant to CPE."

Selected Notes:

CPE.md

MEM.md

 PASS 2: GROUNDED ANSWER SYNTHESIS

To isolate Carbapenemase-Producing Enterobacterales (CPE) strains, the notes mention that the enrichment process has been simplified when correlating with the 'Received: 30 March 2021 microbiota' and facilitating the separation of putative colonies [\[CPE\]](#).

Additionally, a selective enrichment medium, MEM, is used for isolating Coli from Caeca, utilizing selective agents such as cefotaxime (CTX at 1 mg/L) and meropenem [\[MEM\]](#). This selective combination of antibiotics and media components is effective at isolating target Coli under microaerophilic conditions, and established protocols provide a reproducible method for isolating and confirming the presence of target organisms [\[MEM\]](#).

RETRIEVED VAULT CONTEXT (2 NOTES):

CPE.md

CPE > **Executive Summary**: Carbapenemase-Producing...

MEM.md

MEM > **Executive Summary**: MEM is a selective enrichment...

