Università di Genova
DITEN DIPARTIMENTO
DI INGEGNERIA NAVALE, ELETTRICA,
ELETTRONICA E DELLE TELECOMUNICAZIONI

Efficient Compression and Implementation of Domain-Specific Language Models on Resource-Constrained Embedded Systems

A Case Study of a Washing Machine Assistant

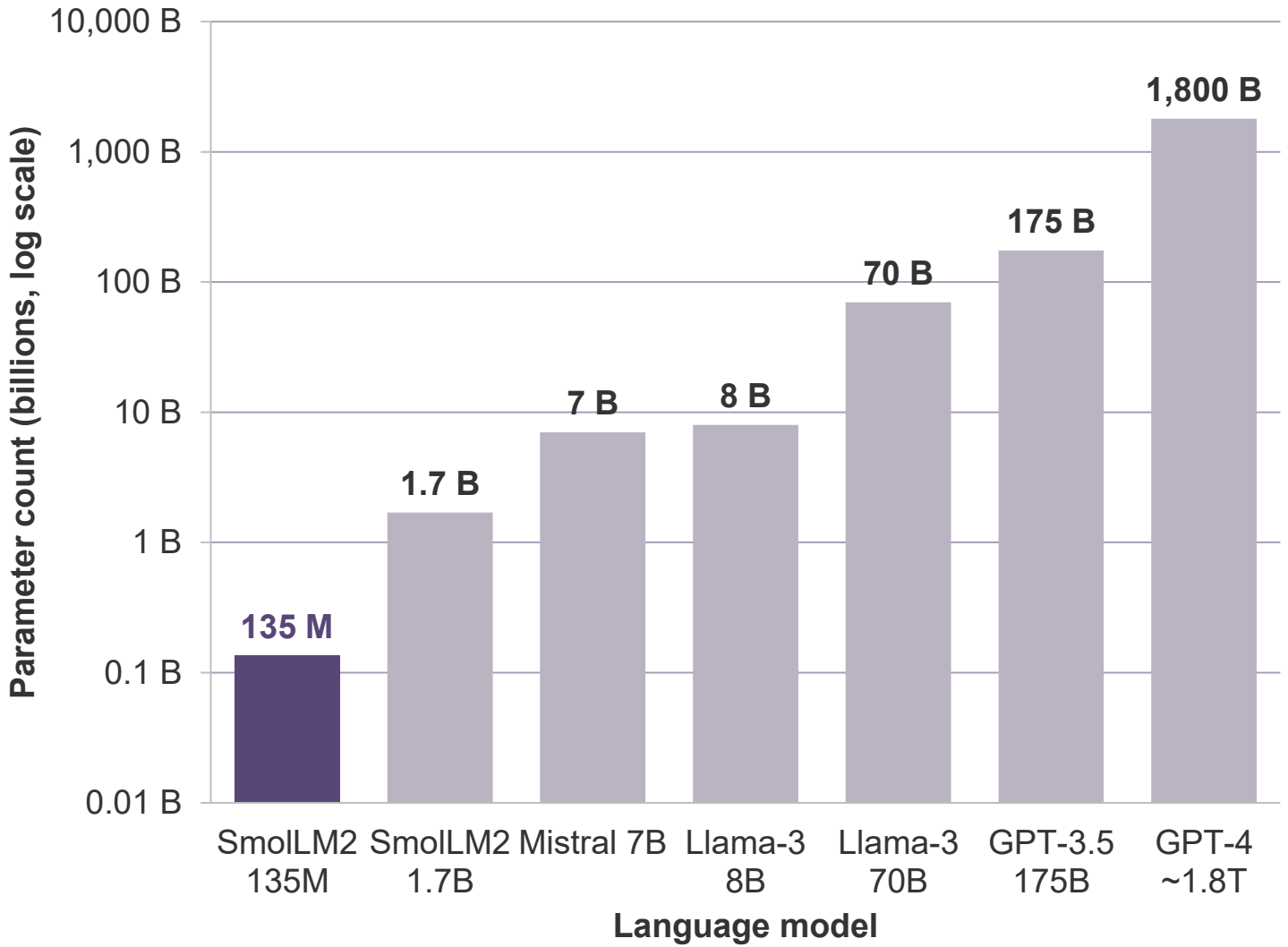
Amin Khakpour

Research Objective

- Create an LLM-based system implemented on a commercial high-end microcontroller.
- Deliver acceptable performance as a specialized chatbot for natural user interaction with common household appliances.
- Case study of a washing machine.

Challenges of deploying LLMs on embedded systems

1. Memory footprint
2. Representational Capacity
3. Absence of hardware acceleration



Targeted hardware: STM32MP257

Specification	Detail
Main CPU	Dual-core Arm Cortex-A35 @ up to 1.5 GHz
Co-processor	Arm Cortex-M33 @ up to 400 MHz
External RAM	Up to 4GB DDR4
Internal SRAM	808 Kbytes

STM32MP257

The plan

Part 1 · Distillation, reasoning transferred from the 1.7B teacher into the 135M student

Part 2 · Dataset engineering

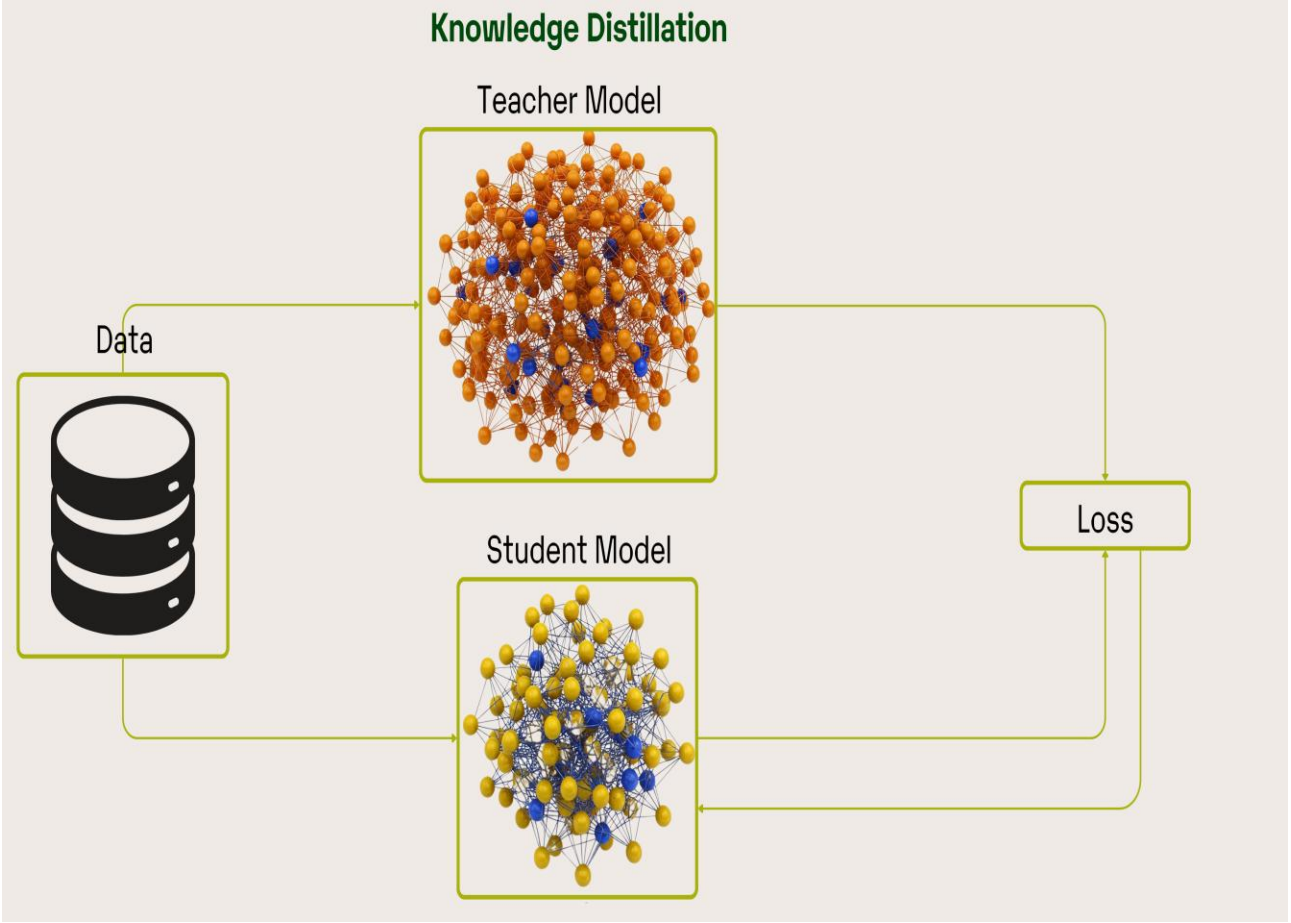
Part 3 · Supervised Fine-Tuning (**SFT**)+ Low-Rank Adaptation(**LoRA**)

Part 4 · Retrieval-Augmented Generation (**RAG**)

Part 5 · GGML Universal Format (**GGUF**) **+ llama.cpp**

Part 1: Knowledge distillation

- Model compression optimization framework.
- Utilizes a massive *Teacher* model.
- Transfers *dark knowledge*.
- Creates a smarter, small *Student* model.



Part 1: Distillation attempts

Stage	Teacher	Student	Outcome
1	Llama-3.2-1B-Instruct	SmolLM2-135M	NaN divergence after the first forward pass. Aligning the vocabulary from 49,152 to 128,256 tokens introduced uncalibrated weights
2	SmolLM2-1.7B-Instruct	SmolLM2-135M	Shared vocabulary, teacher capacity matched to student

Part 1: Distillation evaluation (HellaSwag, **50 MCQ**)

Model Variant	Strategy / Architecture	MC Accuracy	Diagnostic Result
SmolLM2-1.7B (Teacher)	Reference teacher	66.0%	Upper bound
SmolLM2-135M (Baseline)	No distillation	50.0%	Untrained baseline
SmolLM2-135M (Distilled)	400 Samples / 3 Epochs / Ground Truth	58.0%	Success

Part 2: Domain-specific dataset engineering

Function	Role
`set_mode`	Configures the wash profile and target thermal state (mode_name, temperature)
`set_spin_speed`	Adjusts the centrifuge rotation limit for the motor control timer
`stop_cycle`	Triggers an immediate hardware interrupt for safety shutdowns
`start_cycle`	Initiates the active operational phase of the configured wash cycle
`get_status`	Queries the high-level operational phase state machine

The dataset contains 500 samples generated with Gemini 1.5 Pro

Part 3: Supervised fine-tuning

Low-Rank Adaptation (LoRA)

LoRA freezes the original model weights and injects small trainable matrices into the transformer layers. Instead of updating all 135M parameters, only a tiny fraction is trained.

Parameter	Value	Rationale
Rank (r)	16	Capacity sufficient for the domain task within the 4 GB memory budget
Alpha	32	Gives an update scaling of α/r = 2, the ratio used in the original LoRA work
Target modules	Q, K, V, O	Adapters applied to the attention projections only; MLP blocks stay frozen
Epochs	3	Longer schedules overfit the 500-sample set
Learning rate	2e-4	Standard range for LoRA adapters, which update far fewer parameters than full fine-tuning

Part 3: Example SFT outputs

User query	SFT model output	Observed failure
What spin speed should I use for delicate lingerie items?	*“For delicate lingerie items, I recommend using a spin speed of 15-20 revolutions per minute (rpm). … However, if you’re looking for a more precise measurement, you can use a spin speed of 20-30 rps.”*	15-20 rpm and/ 20-30 rps. *15-20 revolutions.*
I need to wash delicate silk shirts. Which washing machine setting should I use and what is the temperature?	*“The most common washing machine setting for delicate silk shirts is the ‘Dirty’ setting. … The temperature of the washing machine setting is usually set at 100 degrees Fahrenheit (38 degrees Celsius). … you’ll need to rinse it with cold water and then rinse it again with hot water. Then, you can rinse it again with cold water and then rinse it again with hot water. Finally, you can rinse it again with cold …”*	No “Dirty” setting exists. 38 °C contradicts the 30 °C specified for silk.

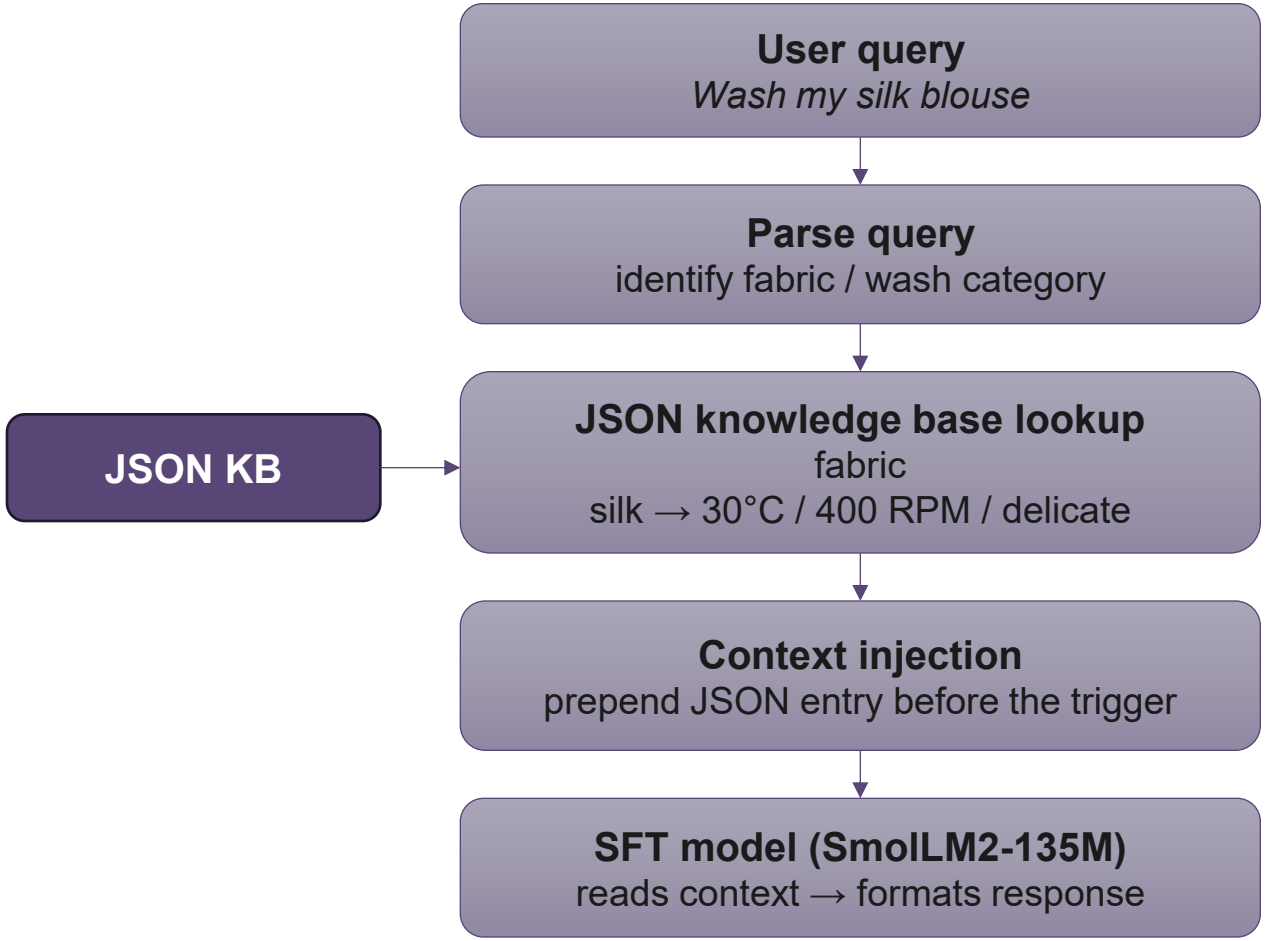
Part 4: RAG implementation

- JSON knowledge base supplies the values for fabric present in the file
- Facts supplied at runtime

```
{
  "fabrics": {
    "silk": {
      "temperature_c": 30,
      "spin_rpm": 400,
      "cycle": "delicate",
      "notes": "Low heat"
    },
    "cotton": { … },
    "wool": { … },
    "synthetics": { … }
  }
}
```

Part 4: How retrieval works

When the query matches a fabric key, the JSON entry is injected into the context and the answer is grounded in it.



Part 4: Evaluation across the four phases

Phase	Faithfulness ↑	Hallucination ↓	Relevancy ↑	Correctness ↑
Base	0.003 ±0.026	0.820 ±0.141	0.334 ±0.056	0.018 ±0.035
Distilled	0.003 ±0.026	0.818 ±0.142	0.333 ±0.066	0.018 ±0.035
SFT	0.004 ±0.026	0.810 ±0.137	0.322 ±0.073	0.018 ±0.034
RAG	**0.304 ±0.224**	**0.590 ±0.190**	**0.414 ±0.176**	**0.301 ±0.231**

Part 5: Deployment toolchain and measurements

- The board can not run the trained model. The target board (STM32MP257) runs Linux on ARM.
- GGUF packages the model, tokenizer, and metadata into one file.
- llama.cpp is a C++ runtime that reads GGUF files and runs on ARM processors.

Model & storage	
Model Size (FP16)	**270 MB**
Model file format	**GGUF, single file**
Inference performance	
Time to first token	**~5.7 s**
Generation speed	**4.91 tok/s**
System resources	
RAM available	**3,771 MB**
RAM used	**519 MB (14%)**
CPU temperature	**~39 °C**

Conclusion

- High memory footprint reduction
- Keeping acceptable accuracy
- Effectiveness of the studied techniques, particularly RAG

Future work

- Profile power consumption and energy efficiency on the target board during inference
- Train on stronger hardware to support a larger student and longer context
- Deploy smaller models on microcontrollers
- Agentic multi-step task execution

Special thanks to

Prof. Agostino

Prof. Bellotti

Prof. Berta

And everyone who helped me through this path

UniGe

DITEN