



Konferencija Srce DEI

# Developing an Accounting Virtual Assistant Through Supervised Fine-Tuning (SFT) of a Small Language Model (SLM)

An evaluation framework for post-trained bookkeeping language models: Preparing for Vibe Bookkeeping

dr.sc. Mario Župan

Sveučilište Josipa Jurja Strossmayera u Osijeku, Fakultet turizma i ruralnog razvoja u Požegi

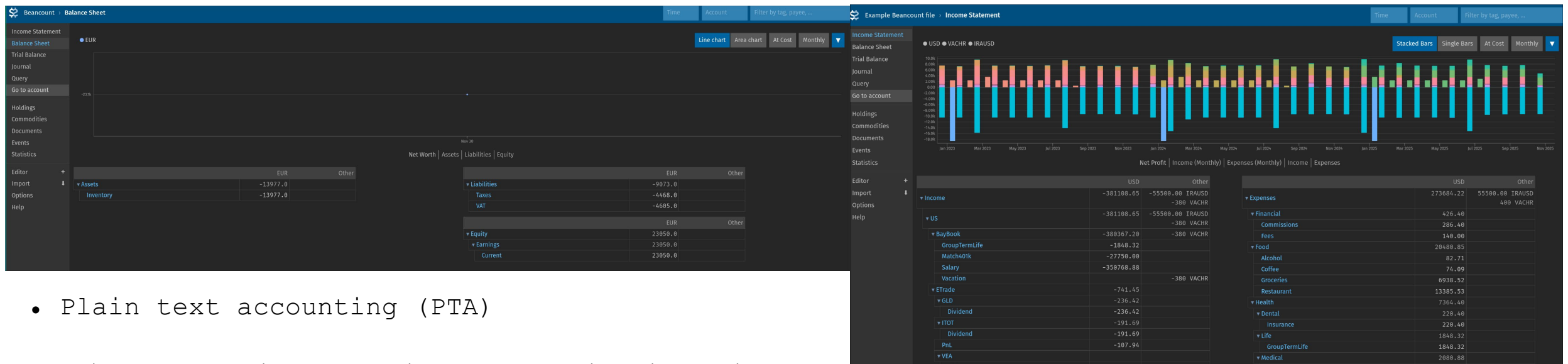
Srce DEI 2026





# The Aim

- Something that is already possible with Large Language models. Well, almost.



- Plain text accounting (PTA)

- is a way of doing bookkeeping and accounting with plain text files and scriptable, command-line-friendly software, such as Ledger, hledger, or Beancount.

- Beancount
- Fava



Something like this: Post a receipt of retail goods into a warehouse, date: ..., money amount: ....

```
2. Add the transaction.
3. Validate the transaction.

We'll use the beancount-subagent to do these.

Let's do it.

# Todos

[.] Provide RETINPUTPKV posting scheme
[ ] Create a corrected ledger file with proper account opening dates and currencies
[ ] Add the RETINPUTPKV transaction to the corrected ledger
[ ] Validate the corrected transaction

Thinking: I need to create a corrected ledger file. First, I should create a new file with proper account opening
dates before the transaction date (2025-11-10) and ensure all accounts use EUR currency. Let me start by creating a
new file.

| Task Create a new beancount file with proper setup
| 1 toolcalls · 13.7s

ctrl+x down view subagents

Thinking: Now I need to add the RETINPUTPKV transaction to this corrected file.

: Task Add RETINPUTPKV transaction to the corrected ledger

ctrl+x down view subagents
```

```
17,953 tokens
2% used
$0.00 spent

MCP
• beancount-mcp Connected
• hf-mcp Connected

LSP
LSPs will activate as files are read

▼ Todo
[.] Provide RETINPUTPKV posting
scheme
[ ] Create a corrected ledger file
with proper account opening
dates and currencies
[ ] Add the RETINPUTPKV transaction
to the corrected ledger
[ ] Validate the corrected
transaction

Let me analyze this request:

1. The user wants a double-entry bookkeeping posting scheme example for document type
"INVOICESB"
2. Based on my training, I need to recall the appropriate accounting standards and
conventions

For document type "INVOICESB", this appears to be for wholesale goods transactions in
Croatian accounting (based on the context). The standard double-entry posting would be:

Debit BLAGAJNA PRODAVAONICE (journal code i.e. account: 1022): 295.68
Credit OBVEZE PDV-A 23% (journal code i.e. account: 24010): 55.43
Debit ROBA U PRODAVAONICI (journal code i.e. account: 6630): 0.00
Credit UKALKULIRANI PDV U PRODAJNU CIJENU (journal code i.e. account: 664): 0.00
Credit RAZLIKA U CIJENI ROBE U MALOPRODAJI (journal code i.e. account: 6690): 0.00
Credit PRIHODI OD PRODAJE ROBE NA MALO (journal code i.e. account: 7610): 240.25

This follows standard accounting practices where:
- Debits = Credits = 280.07
- We're recognizing revenue from goods sold
- We're properly allocating VAT obligations
- We're handling inventory valuation differences

The journal codes and account names follow standard Croatian accounting conventions.

~ Loading skill...

■ Knjigovodja · seed-oss-default-config-q8_0.gguf
```

```
Let me analyze the requirements.

Context
0 tokens
0% used
$0.00 spent

MCP
• beancount-mcp Disabled
• hf-mcp Disabled

LSP
LSPs will activate as files are read
```



# Resources

- Security issues with LLMs -> confidential, proprietary data!
- Current state:
  - Structured dataset <50K rows
    - Double-entry bookkeeping posting scheme examples
  - Equipment:
    - GPUs on gpu-bigmem node: 4x96GB PG506-232
    - mem = 493GB per node



# Possible solutions

- STRATEGY A) Agents.md BUT: AGENTS.md instructions cannot be too large > limited with context window size of the LLM
- STRATEGY B) Hugging Face MCP Server and/or semantic search
  - constraints: bloating the context window!
- STRATEGY C) Posttraining, Supervised finetuning, Learning for specific task
  - QLoRA on 20B-100B LMs
    - ~2-5h training;
    - ~2-13h evaluation;
- STRATEGY D) FFT on ~20B LMs



# Strategies C) and D)

- LMs:
  - Reasoning vs Non-reasoning models
    - (LLM's that return their CoT inside a key called reasoning\_content and then you have other LLM's that return their CoT between <think> ... </think> tags.)
  - MoE vs Dense architecture
    - speed
    - quality
  - Instruct vs Base models
    - Agent calling



# Steps

- Creating chat template dataset(s)
- Preprocesses and tokenizes datasets
- Selecting a LM by:
  - quality
  - size you are capable to train
- Train
- Merge
- Convert to gguf (llama.cpp)
- Quantize to q8\_0
- Evaluate
- Inference on gpu-bigmem

```
[
  {
    "content": "('Provide a double-entry bookkeeping posting scheme example for The document RETOUTPUT (Document codename for posting invoices for retail goods sold and for moving goods from retail stock. Revenue and VAT liabilities are included, and the invoice is generated by the cash register.). It specified a business process based on Croatian tax and accounting regulations for the year 2009.').",
    "role": "user"
  },
  {
    "content": "{\"year\": \"2009\", \"document\": \"RETOUTPUT\", \"document_explanation\": \"Document codename for posting invoices for retail goods sold and for moving goods from retail stock. Revenue and VAT liabilities are included, and the invoice is generated by the cash register. \\\", \"total_debit\": 13082.35, \"total_credit\": 13082.35, \"entries\": [{\"ACCOUNT\": \"1022\", \"TITLE\": \"BLA\", \"DEBIT\": 15464.79, \"CREDIT\": 0.0}, {\"ACCOUNT\": \"24010\", \"TITLE\": \"OBV\", \"DEBIT\": 0.0, \"CREDIT\": 2788.73}, {\"ACCOUNT\": \"6630\", \"TITLE\": \"ROB\", \"DEBIT\": -2382.44, \"CREDIT\": 0.0}, {\"ACCOUNT\": \"664\", \"TITLE\": \"UKA\", \"DEBIT\": 0.0, \"CREDIT\": -429.62}, {\"ACCOUNT\": \"6690\", \"TITLE\": \"RAZ\", \"DEBIT\": 0.0, \"CREDIT\": -1952.82}, {\"ACCOUNT\": \"7610\", \"TITLE\": \"PRI\", \"DEBIT\": 0.0, \"CREDIT\": 12676.06}]}\",
    "role": "assistant"
  }
]
```

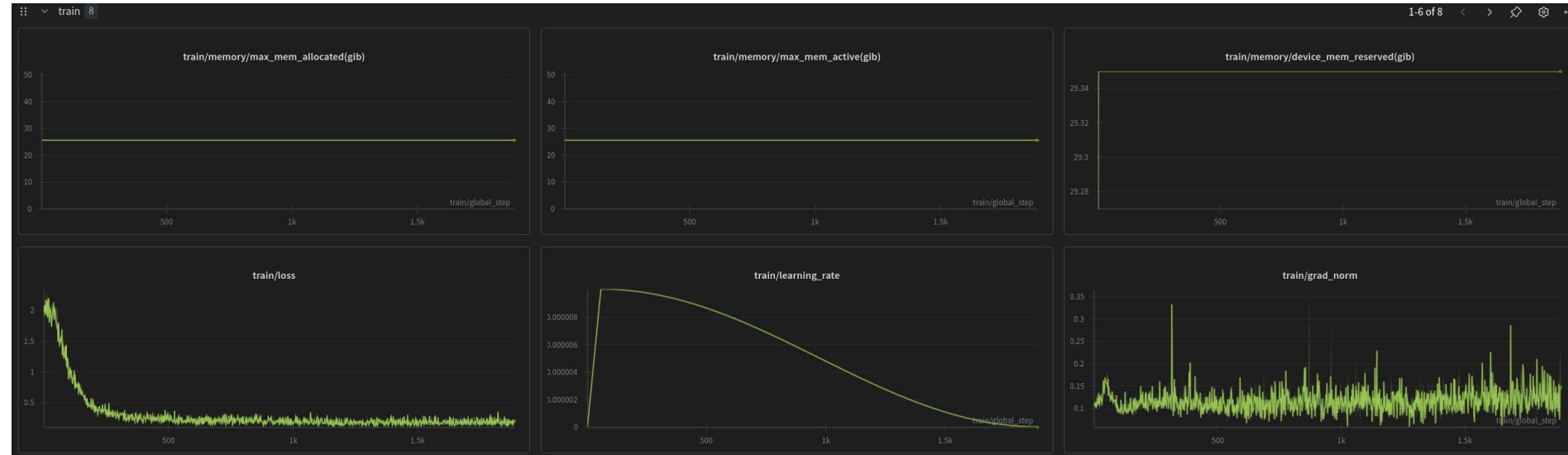
Model	Total Parameters	Active Parameters	Architecture	Key Strengths / Notes
GPT-OSS-20B	~20B	~20B	Dense Transformer	High factual knowledge (MMLU 90.0), exceptional scientific QA (GPQA-D 80.9), top instruction following (IFEval 92.8)
GLM-4.5-Air	106B	12B	Sparse MoE Transformer	Hybrid reasoning (thinking/non-thinking modes), optimized for agent tasks, strong efficiency for its active size
Seed-OSS-36B-Instruct	36B	36B	Dense Transformer	Efficiently trained on only 12T tokens; outstanding reasoning depth (MMLU-Pro 82.7, BBH 87.7) and math (GSM8K 90.8)



# Training

Weak training:

- num\_epochs: 1-2
- learning\_rate: 2e-5
- lora\_r: 32



```
{'loss': 1.9609, 'grad_norm': 0.11034293472766876, 'learning_rate': 0.0, 'memory/max_mem_active(gib)': 25.68, 'memory/max_mem_allocated(gib)': 25.68,
'memory/device_mem_reserved(gib)': 29.27, 'epoch': 0.0}
2026-01-22 09:58:31
...
{'loss': 0.1766, 'grad_norm': 0.1368938535451889, 'learning_rate': 6.623759873924939e-11, 'memory/max_mem_active(gib)': 25.68, 'memory/max_mem_allocated(gib)': 25.68,
'memory/device_mem_reserved(gib)': 29.35, 'epoch': 0.3}
2026-01-22 14:40:37
{'loss': 0.1147, 'grad_norm': 0.13559143245220184, 'learning_rate': 2.943897209339852e-11, 'memory/max_mem_active(gib)': 25.68, 'memory/max_mem_allocated(gib)': 25.68,
'memory/device_mem_reserved(gib)': 29.35, 'epoch': 0.3}
...
{'train_runtime': 17000.0526, 'train_samples_per_second': 0.888, 'train_steps_per_second': 0.111, 'train_loss': 0.3049940504469313, 'memory/max_mem_active(gib)': 25.68,
'memory/max_mem_allocated(gib)': 25.68, 'memory/device_mem_reserved(gib)': 29.35, 'epoch': 0.3}
...
```



# Evaluation: Gpt-oss-fft-1-epoch

## GPT-OSS-20B Full Finetuning (1 Epoch) - Summary

### 📊 Bookkeeping Task Performance (NEW KNOWLEDGE)

- **Parsed (JSON validity):** 1.000 ✓ Perfect
- **Balanced (debit=credit):** 0.955 ✓ Excellent
- **Accounts Correct:** 0.518 ✓ Strong
- **Accounts Score:** 0.736 ✓ Best among all models evaluated
- **Balanced+Score:** 0.705 ✓ Top performance

### 📊 General Knowledge Performance (PRESERVATION ASSESSMENT)

Compared to expected base model performance for similar-sized LLMs:

Benchmark	Score	Expected Base	Change	Assessment
Hellaswag (acc_norm)	0.453	~0.65-0.75	-30% to -40%	Significant degradation
MMLU (overall acc)	0.229	~0.35-0.45	-40% to -50%	Major knowledge loss
ARC-Easy (acc)	0.410	~0.55-0.65	-30% to -40%	Substantial decline
GSM8K (flexible-extract)	0.453	~0.10-0.15	+200% to +350%	Large improvement
Winogrande (acc)	0.512	~0.65-0.75	-30% to -40%	Significant degradation
TruthfulQA (mc2)	0.564	~0.50-0.55	~0% to +10%	Similar/Slight gain
IFEval (inst_level_strict_acc)	0.255	~0.45-0.55	-50% to -55%	Severe instruction following loss



# Evaluation: Seed base vs Seed QLoRA

Benchmark	Base Score	Finetuned Score	Change	Explanation
hellaswag	0.683	0.740	+0.057	Common sense reasoning
mmLu	0.230	0.514	+0.284	Multi-task understanding (57 tasks)
truthfulqa	0.356	0.357	+0.001	Truthfulness/honesty
winogrande	0.684	0.707	+0.023	Commonsense reasoning
arc	0.428	0.524	+0.096	Abstract reasoning
gsm8k	0.115	0.289	+0.174	Math word problems
ifeval	0.474	0.450	-0.024	Instruction following

## Key Findings

- New knowledge acquired:** Strong performance on bookkeeping task confirms domain adaptation
- General knowledge preserved:** All major reasoning benchmarks show improvement
- Significant gains:** Particularly notable in mathematical reasoning (GSM8K +151%) and broad knowledge (MMLU +123%)
- Minimal trade-off:** Only slight decrease in IFEval instruction following (-2.4%), while other metrics improved substantially

The results indicate the QLoRA finetuning process successfully added new accounting knowledge without degrading - and actually enhancing - the model's general capabilities. This suggests the finetuning data and methodology provided beneficial regularization or knowledge transfer effects.



# Evaluation

Comprehensive Model Comparison: New Knowledge Acquisition vs. General Knowledge Preservation

Model	Architecture	Finetuning Method	Bookkeeping Performance	General Knowledge (MMLU)	General Knowledge Trend	Common Capability
<b>GPT-OSS-20B-FFT</b>	Dense	Full (1 epoch)	<b>Best:</b> Parsed=1.000, Accounts Correct=0.518, Accounts Score=0.736	0.229	<b>Severe Degradation</b> (Est. -40-50% from base)	Designed for multi-step reasoning, problem solving, and structured logic generation
<b>Seed-OSS-36B-QLoRA</b>	Dense	QLoRA	<b>Strong:</b> Parsed=0.993, Accounts Correct=0.424, Accounts Score=0.731	0.514	<b>Improvement</b> (+28.4% from base 0.230)	Designed for multi-step reasoning, problem solving, and structured logic generation
<b>GLM-4-5-Air-QLoRA</b>	MoE	QLoRA	<b>Moderate:</b> Parsed=0.918, Accounts Correct=0.087, Accounts Score=0.406	0.331	<b>Improvement</b> (Est. +10-15% from base)	Designed for multi-step reasoning, problem solving, and structured logic generation



# Evaluation

## Explanation:

- **Parsed**: Ability to generate valid JSON output (required format)
- **Balanced**: Correctly maintaining debit=credit equality (fundamental accounting principle)
- **Accounts Correct**: Exact match of required account selections
- **Accounts Score**: Partial credit scoring for account selection quality
- **Balanced+Accounts**: Combined metric requiring both balancing AND correct accounts

## 🔍 Detailed Benchmark Explanations & Scores

### Bookkeeping Task (New Knowledge Acquisition)

*Measures ability to learn double-entry accounting from finetuning data*

Model	Parsed (JSON Validity)	Balanced (Debit=Credit)	Accounts Correct	Accounts Score	Balanced+Accounts
GPT-OSS-20B-FFT	1.000	0.955	<b>0.518</b>	0.736	<b>0.503</b>
Seed-OSS-36B-QLoRA	0.993	0.990	0.424	<b>0.731</b>	0.424
GLM-4-5-Air-QLoRA	0.918	0.832	0.087	0.406	0.084



# Evaluation

## General Knowledge Benchmarks (Knowledge Preservation Assessment)

Measures retention of broad capabilities after finetuning

Benchmark	Explanation	GPT-OSS-20B-FFT	Seed-OSS-36B-QLoRA	GLM-4-5-Air-QLoRA
Hellaswag	Common sense reasoning	0.453 (acc_norm)	0.740	0.708
MMLU	Multi-task understanding (57 tasks)	0.229	<b>0.514</b>	0.331
TruthfulQA	Truthfulness/honesty (resists misinformation)	0.564 (mc2)	0.548	0.535
Winogrande	Commonsense reasoning / word understanding	0.512	0.707	0.718
ARC	Abstract reasoning (challenging science puzzles)	0.410 (acc_norm)	0.524	0.438
GSM8K	Math word problem reasoning	<b>0.453</b> (flexible)	0.289	0.396
IFEval	Instruction following / prompt adherence	0.255 (inst_level_strict)	0.450	0.459



# Conclusion

QLoRA-finetuned Seed-OSS-36B model demonstrates both successful acquisition of new knowledge and preservation/enhancement of general knowledge:

## New Knowledge Acquisition (Double-Entry Accounting)

The model shows strong performance on the `bookkeeping_posting_schemes_validity` task:

- **Parsed**: 0.993 (near-perfect JSON generation)
- **Balanced**: 0.990 (near-perfect debit/credit balancing)
- **Accounts Score**: 0.731 (good partial credit scoring)
- **Accounts Correct**: 0.424 (moderate account selection accuracy)

This confirms the model successfully learned the double-entry accounting domain from the finetuning data.

- The model successfully learned the double-entry accounting domain from the finetuning data.
- General Knowledge Preservation & Enhancement
- No catastrophic forgetting
- The model shows significant improvements across general knowledge benchmarks



# Conclusion

## Key Insights

### Best Overall Balance: Seed-OSS-36B-QLoRA

- **Strong bookkeeping performance** (near-top in all metrics)
- **Significant general knowledge IMPROVEMENT** (+28.4% MMLU)
- **Only model showing enhancement** across both dimensions
- Suggests QLoRA on dense architecture may provide beneficial regularization

### Specialization Champion: GPT-OSS-20B-FFT

- **Absolute best bookkeeping performance** (perfect JSON, highest account selection)
- **Severe general knowledge degradation** (estimated 40-50% losses)
- Demonstrates **catastrophic forgetting** with full parameter updates
- Only major win in GSM8K (math reasoning transfer)

### Moderate Performer: GLM-4-5-Air-QLoRA

- **MoE architecture** (confirmed via parameters=106B)
- **Reasonable bookkeeping** but weakest account selection (0.087)
- **Modest general knowledge gains** (+10-15% MMLU estimated)
- Shows MoE can learn new domains with less forgetting than full FT



# Conclusion

## 🎯 Critical Findings

1. **Architecture Matters:** Dense models (Seed, GPT-OSS) adapted to bookkeeping more effectively than MoE (GLM) in absolute terms
2. **Method Trumps Scale:** Seed-OSS-36B (QLoRA, 36B) outperformed GPT-OSS-20B (full FT, 20B) in general knowledge despite smaller size
3. **QLoRA Advantage:** Both QLoRA models preserved/enhanced general capabilities while learning new domain
4. **Full Finetuning Risk:** GPT-OSS-20B's catastrophic forgetting demonstrates danger of full parameter updates for domain adaptation
5. **Unexpected Transfer:** GSM8K improvements across all models suggest mathematical reasoning benefits from accounting training

## 💡 Recommendation

For **dual-objective scenarios** requiring both:

- **New domain acquisition** (bookkeeping)
- **General capability preservation/improvement**

**Parameter-efficient methods (QLoRA/LoRA) on dense architectures** (like Seed-OSS-36B) provide optimal balance, avoiding the severe forgetting seen in full finetuning while achieving strong specialized performance. The MoE architecture showed more resistance to forgetting than full FT but less capacity for new knowledge acquisition than dense models with QLoRA.



## 📍 Critical Findings

1. **Architecture Matters:** Dense models (Seed, GPT-OSS) adapted to bookkeeping more effectively than MoE (GLM) in absolute terms
2. **Method Trumps Scale:** Seed-OSS-36B (QLoRA, 36B) outperformed GPT-OSS-20B (full FT, 20B) in general knowledge despite smaller size
3. **QLoRA Advantage:** Both QLoRA models preserved/enhanced general capabilities while learning new domain
4. **Full Finetuning Risk:** GPT-OSS-20B's catastrophic forgetting demonstrates danger of full parameter updates for domain adaptation
5. **Unexpected Transfer:** GSM8K improvements across all models suggest mathematical reasoning benefits from accounting training

## 💡 Recommendation

For **dual-objective scenarios** requiring both:

- **New domain acquisition** (bookkeeping)
- **General capability preservation/improvement**

**Parameter-efficient methods (QLoRA/LoRA) on dense architectures** (like Seed-OSS-36B) provide optimal balance, avoiding the severe forgetting seen in full finetuning while achieving strong specialized performance. The MoE architecture showed more resistance to forgetting than full FT but less capacity for new knowledge acquisition than dense models with QLoRA.



Ovo djelo je dato na korištenje pod licencom Creative Commons *Imenovanje* 4.0 međunarodna.

Srce politikom otvorenog pristupa široj javnosti osigurava dostupnost i korištenje svih rezultata rada Srca, a prvenstveno obrazovnih i stručnih informacija i sadržaja nastalih djelovanjem i radom Srca.

[www.srce.unizg.hr](http://www.srce.unizg.hr)

[creativecommons.org/licenses/by/4.0/deed](https://creativecommons.org/licenses/by/4.0/deed)

[www.srce.unizg.hr/otvoreni-pristup](http://www.srce.unizg.hr/otvoreni-pristup)

