



Hochschule Aalen
UNIVERSITY OF APPLIED SCIENCES

Follow-Up Projects

SEMINAR & MASTER-THESIS
TOPICS

Three frontiers you can actually join

CLUSTER 1

LLM Inference & Efficiency

Serve a million-token context without a terabyte of KV cache. Measure the GQA → MLA → sparse frontier.

[builds on Lecture 5](#)

CLUSTER 2

Evaluating Code & Agents

"95% on SWE-bench" measures a *system*, not a model. Take the benchmark apart and calibrate it.

[builds on Lectures 8–10](#)

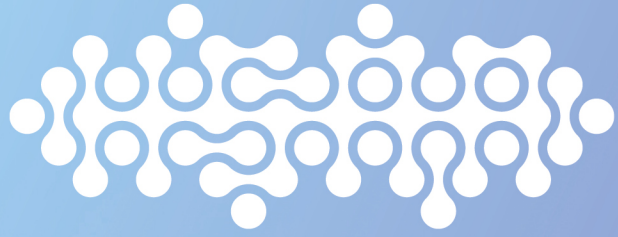
CLUSTER 3

Transformers for Vision

The same attention you know — on image patches. DETR detection & frozen DINOv2 features.

[same transformer, new modality](#)

Each cluster has a 1-page deep-dive with ready-to-start topics. This deck is the trailer.



Hochschule Aalen
UNIVERSITY OF APPLIED SCIENCES

LLM Inference & the Efficiency Frontier

1

The lecture wrote the math. Nobody measured it.



KV cache **per request** · DeepSeek-V3-class model · 32K context (Lecture 5 numbers). Decoding is *memory-bound*: ~128 GB of cache reads *per token*.

The 2026 frontier (DeepSeek-V4; **FlashMemory**, arXiv:2606.09079) reframes the indexer as *a retriever you can train without the backbone* — runnable on one GPU precisely because you never touch the giant model.

Cluster 1 — the project menu

SEMINAR / THESIS ★

P1 · GQA vs MLA vs Sparse — a reproducible single-GPU KV-cache benchmark. Measured vs predicted, with error bars.

SEMINAR / THESIS

P2 · When Top-K Misses the Needle — when does sparse attention silently drop the one token that mattered? RULER + failure taxonomy.

AMBITIOUS

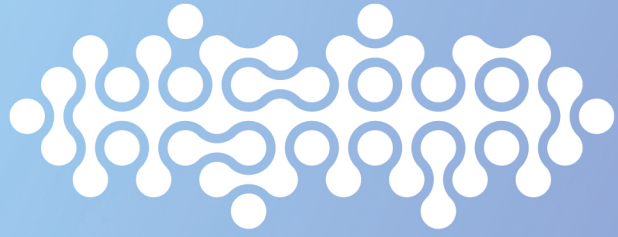
P3 · The Indexer Is a Retriever — a decoupled neural-memory indexer, trained without the backbone. The FlashMemory idea, tested.

SEMINAR / THESIS

P4 · The Efficient-Attention Zoo — a structured map from KV cache to 1M-token serving. One slide, the whole field.

SEMINAR / THESIS · OPT

P5 · Speculative Decoding vs the Memory Wall — does drafting help when you are bandwidth-bound? Composition study.



Hochschule Aalen
UNIVERSITY OF APPLIED SCIENCES

Evaluating Code Generation & Agents

2

"95% on SWE-bench" is a measurement of a *system*

$$95\% = \boxed{\text{Model}} + \boxed{\text{Scaffold}} + \boxed{\text{Grader}} + \boxed{\text{Data}}$$

Change any one of them — retries, test strength, leaked issue threads — and the number moves. Usually those conditions are **undisclosed**.

reported 12.47%

3.97%

← what a manual audit does once you remove leakage + weak tests (SWE-Bench+). On genuinely fresh issues: 0.55%.

Cluster 2 — the project menu

SEMINAR / THESIS

P1 · The Scaffold Tax — fix the model, vary only the harness: how many points are the scaffold? Agentless vs OpenHands.

SEMINAR / THESIS

P2 · Audit the Wins — re-read "resolved" instances by hand; replicate SWE-Bench+. Leakage vs weak tests.

SEMINAR / THESIS ★

P3 · Cost-per-Solved-Task — restore the suppressed cost axis; does the ranking reorder? Accuracy-per-euro Pareto.

SEMINAR / THESIS

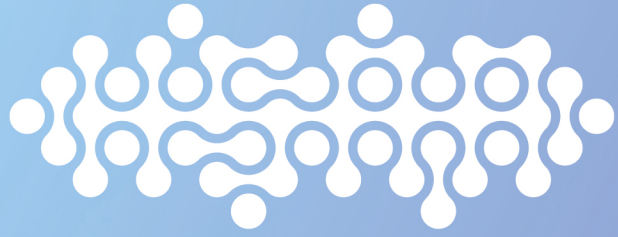
P4 · Does context engineering help? — replicate the AGENTS.md "cost without gain" result. Success AND cost.

SEMINAR / THESIS

P5 · Benchmark-quality meta-eval — score today's coding benchmarks with a rubric. + one cheap diagnostic.

SEMINAR / THESIS

P6 · Faithfulness & hallucination — what should a faithfulness benchmark even measure? Cross-domain TRUST seed.

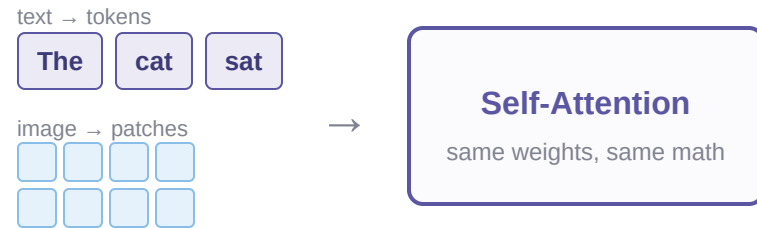


Hochschule Aalen
UNIVERSITY OF APPLIED SCIENCES

Transformers Beyond Text: Vision

3

The same transformer — pointed at a new modality



Detection became set prediction. DETR's object queries are *learned slots*. DETR → RT-DETR: NMS-free, now real-time.

frozen **Features became free.** A DINOv2 backbone + a tiny trainable head — single-GPU, the "freeze the LLM, train a LoRA" move.

Cluster 3 — the project menu

SEMINAR / THESIS

P1 · DETR → Real-Time — map the lineage, then fine-tune RT-DETR vs YOLO on a domain set. Do DETRs finally beat YOLO?

SEMINAR / THESIS ★

P2 · Frozen DINOv2 + Light Head — how far do frozen features get on a domain task? Label-efficiency curve. The flagship.

SEMINAR / THESIS

P3 · Patch-Tokens Under the Microscope — ViT vs CNN + attention explainability. Attention you already know.

SEMINAR / THESIS

P4 · Transferring the LLM Lesson — token-merging efficiency for a ViT. The vision mirror of efficient attention.

AMBITIOUS

P5 · Few-Shot Domain Detector — frozen foundation features + a light detection head. Strong student only.

Pick your frontier

Cluster	Pick if you like...	Prereqs	Safe flagship
1 · Inference	GPUs, systems, the math of serving	KV-Cache 5, PyTorch	Reproducible KV-cache benchmark
2 · Eval & Agents	measurement, skepticism, tooling	Python, Docker, CLI	Cost-per-solved-task Pareto
3 · Vision	images, transfer learning, fast wins	attention lectures, PyTorch	Frozen DINOv2 + light head

Every topic here is single-GPU, public-data, and 6-month-real. The deep-dive for each cluster lists research questions, milestones, an honest scope, and a vetted reading list.

