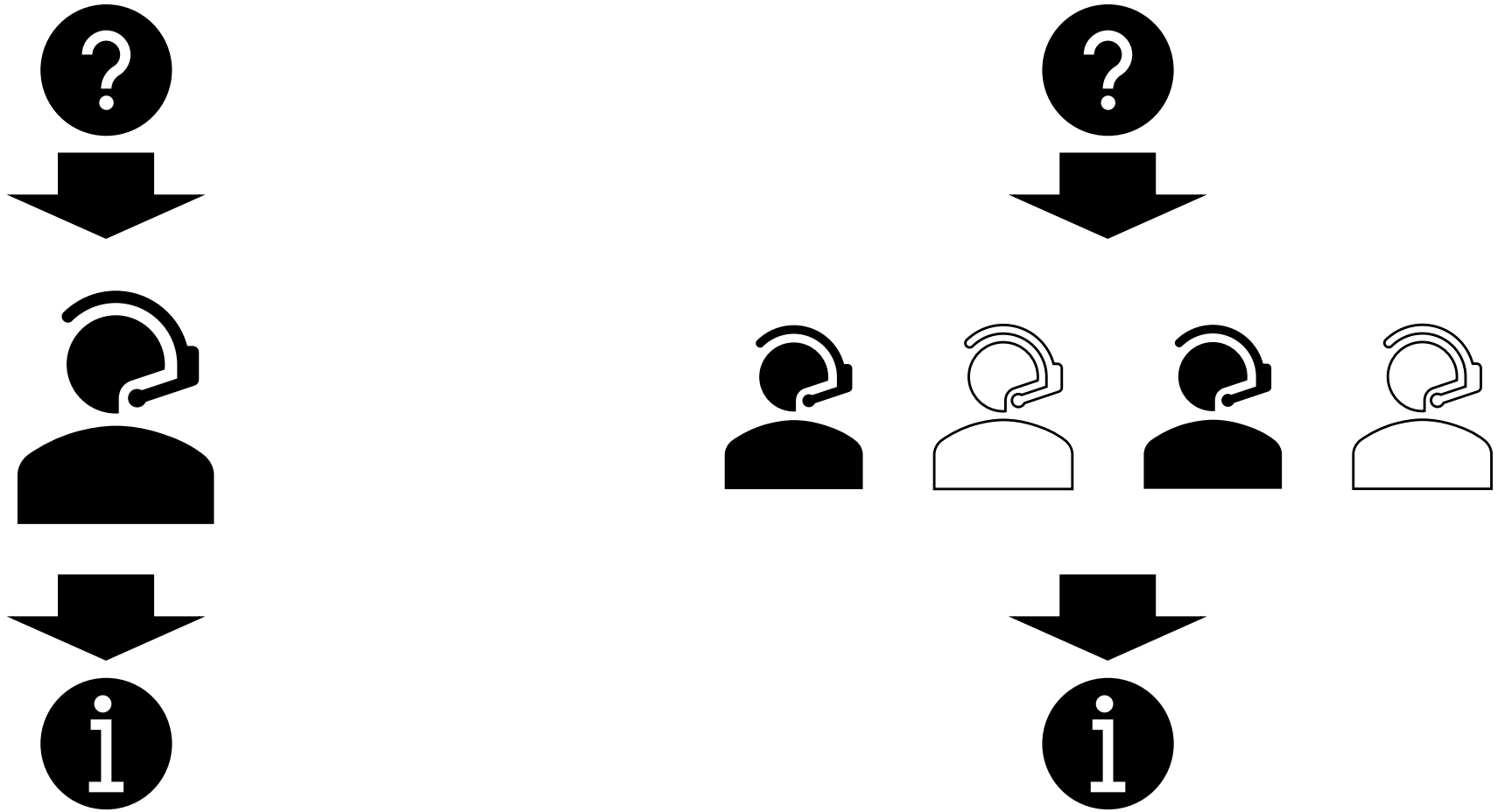
Seminar in NLP (SS26)

Mixture-of-Experts Architectures in Large Language Models

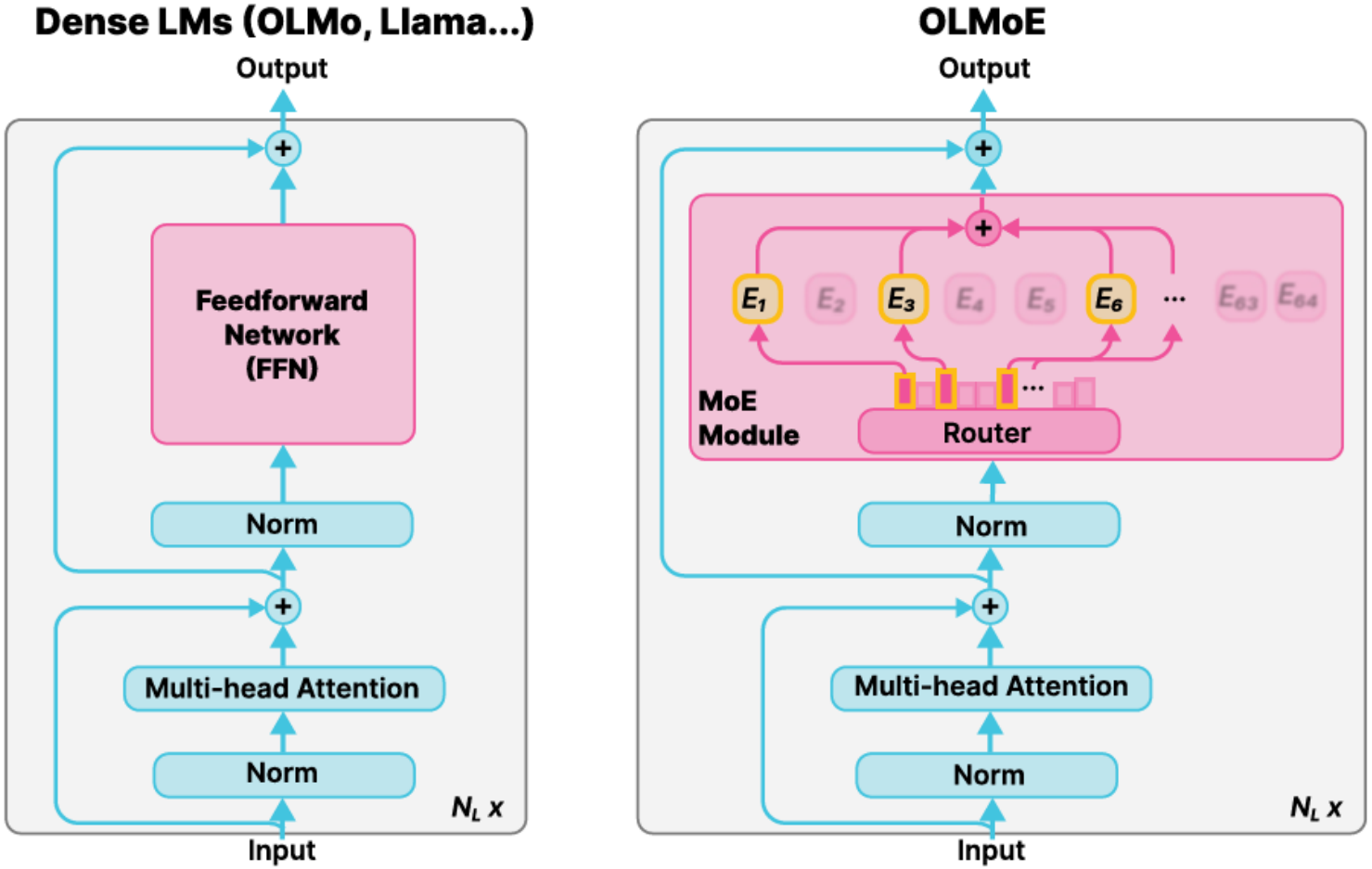
CAIDAS Chair for Natural Language Processing
Benedikt Ebing
Saad Obaid
Prof. Dr. Goran Glavaš

Goal: Write & present a literature survey about your topic

Why should you care about Mixture-of-Experts?



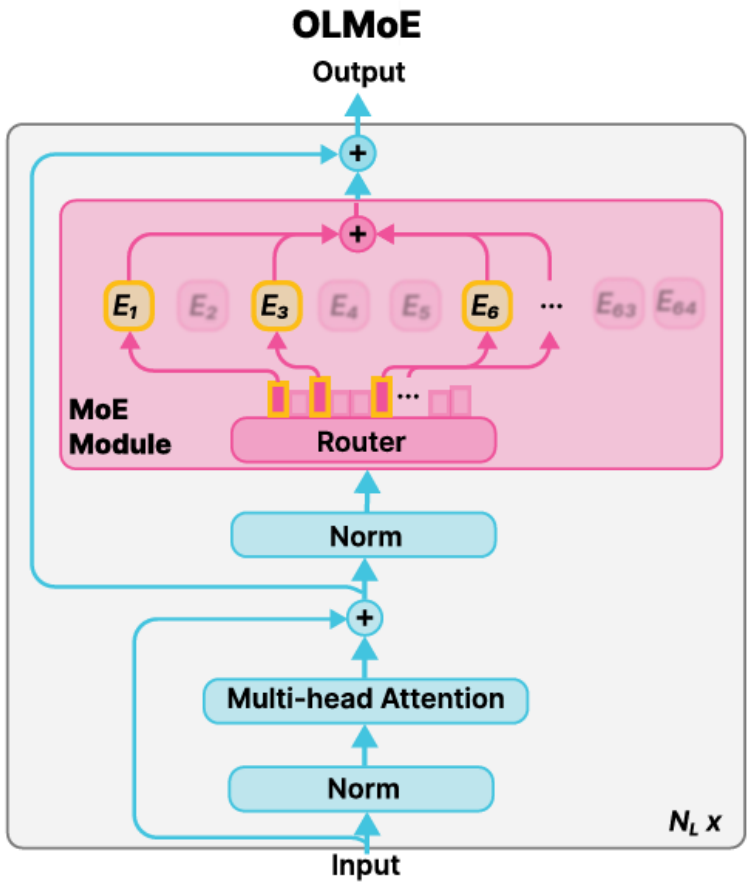
What are MoEs?



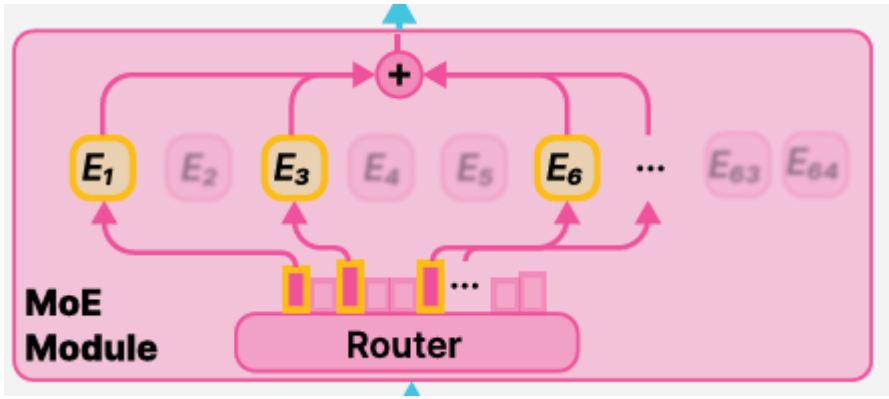
Subtopics

- **Architecture (Expert Design, Routing, Load Balancing)**
- **Upcycling**
- **Post-Training (Instruction Tuning, LoRA Fine-Tuning)**
- **Domain and Language Modularization**

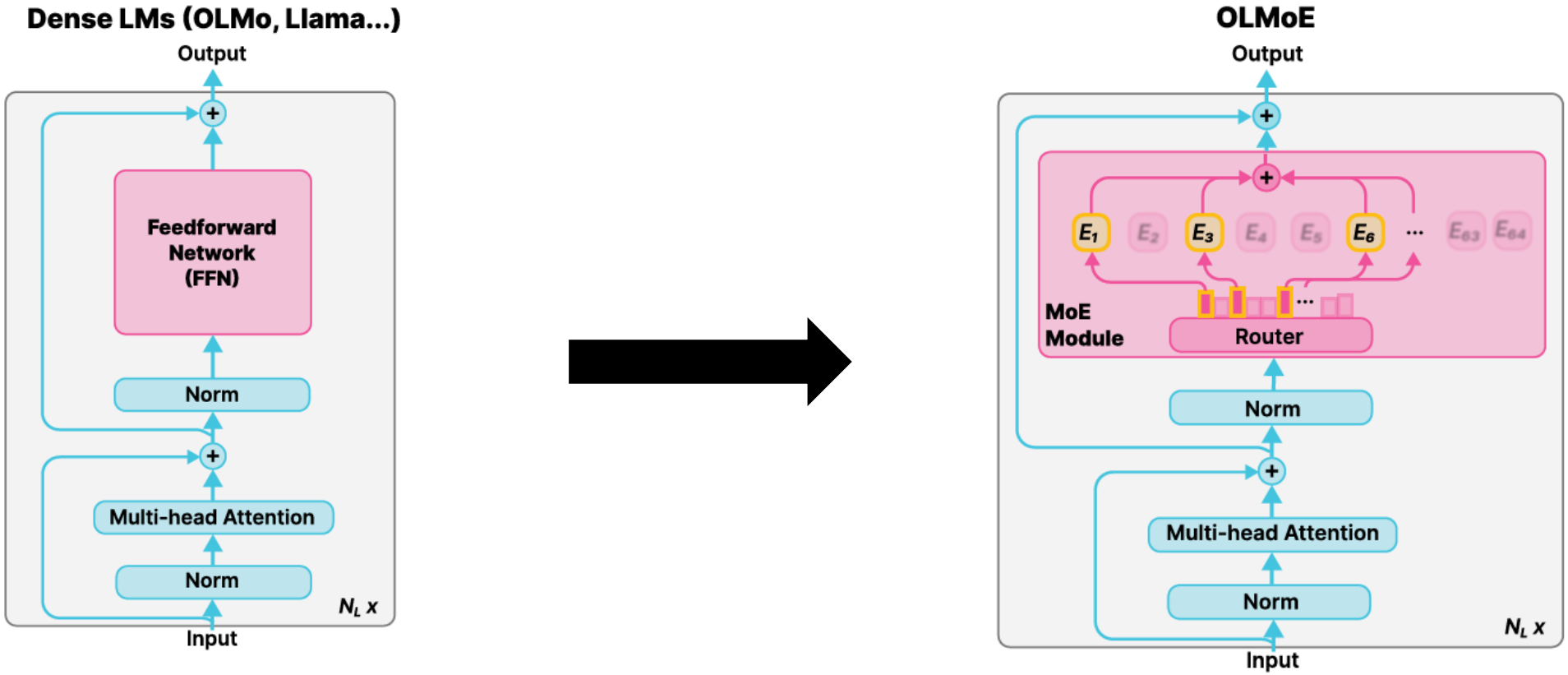
Subtopic: Architecture



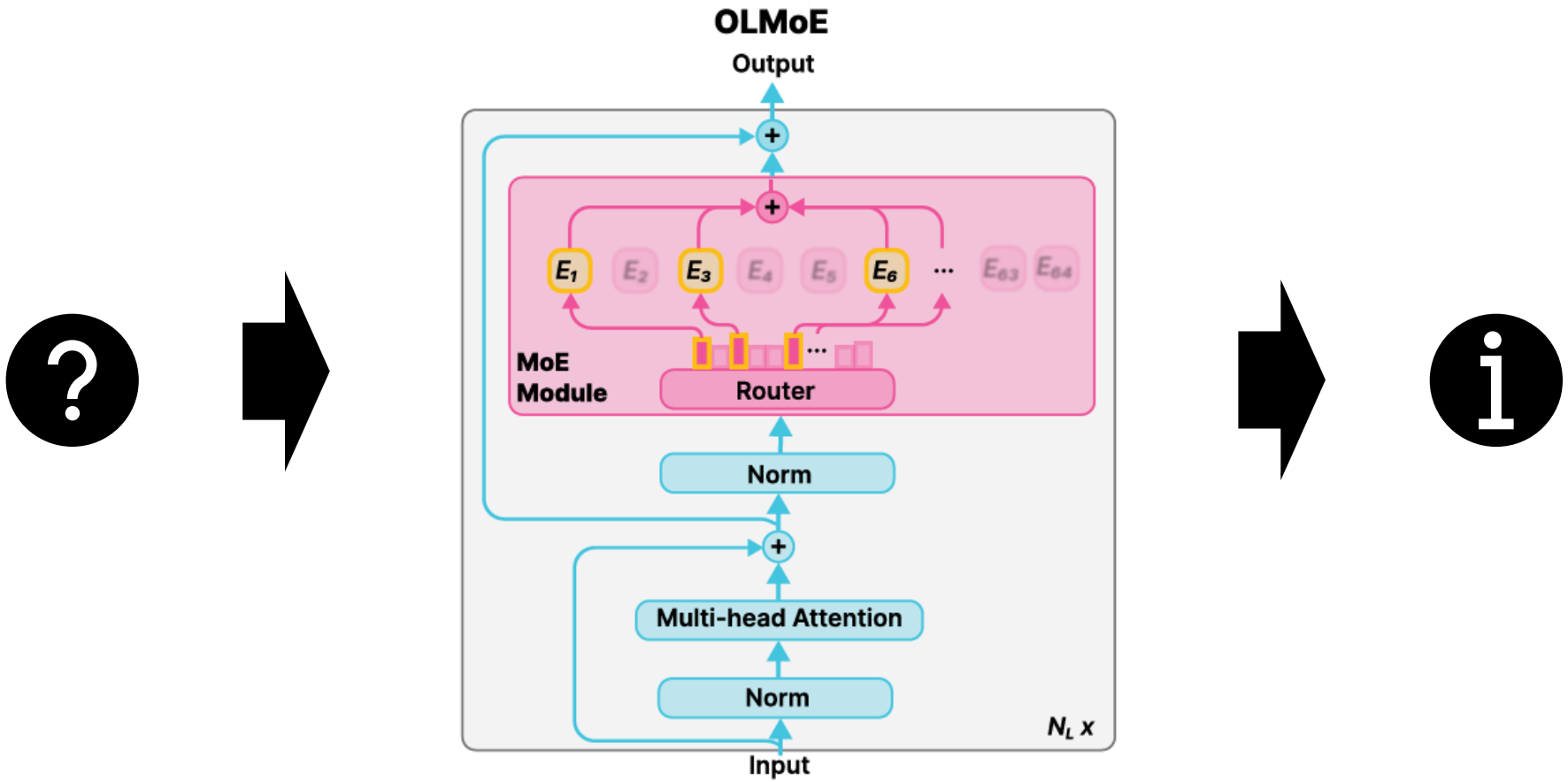
How to design this module?



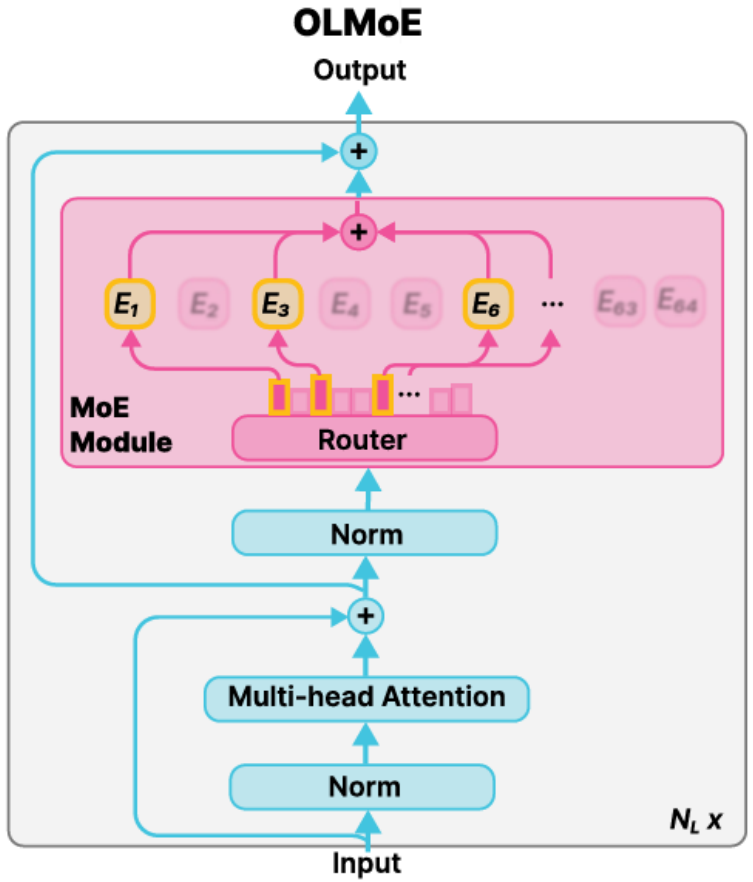
Subtopic: Upcycling



Subtopic: Post-Training



Subtopic: Domain and Language Modularization



Presentation

- 10-15 minutes
- What, why, and how
- 5 minutes Q&A
- Target audience: your fellow students

Report

- Use LaTex template
- 6 – 8 pages
- Use your own words
- Follow good scientifc practice: e.g.,
  - Cite all related work, properly
  - Mark direct citations (if necessary)
- Target audience: as for the presentation

TODO!

Send you topic preference to benedikt.ebing@uni-wuerzburg.de; include your **transcript of records** and a **4-6 sentence paragraph about your motivation** (*Why are you interested?*, *Why are you a good fit?*)

Deadline: this Friday, **24.4.2026**

Acceptance information: next Monday, **27.4.2026**

More details in the WueCampus Course

Feedback

- We offer 2 feedback sessions + on demand feedback via E-Mail
- 1. Session – beginning of June:
  - Present your structured current literature review
  - Show understanding of background concepts and papers
- 2. Session – beginning of July:
  - Shaping your presentation (draft) and the outline of your report
- Come prepared: A few slides summarizing your current progress; clear idea of what you want to discuss

Details / How-To

Seminar Description

- **Goal:** Present and write a survey paper about Mixture-of-Experts Architectures in Large Language Models focused on your subtopic.
- Read, understand and explore scientific literature (with listed papers as starting points for your analysis)
- Organize the collected knowledge for a meaningful presentation about your topic
  - 10-15 minutes + 5 minutes Q&A
- Summarize your topic in a concise report
  - 6 - 8 pages + references

Read, understand and explore scientific literature
➔ From exploration to deep understanding

- Your goal is to understand the **main ideas**, **key results**, and **current research direction** related to your subtopic
  - Why is the topic relevant?
  - What are the problems?
  - How are the problems tackled? (e.g., methods and models)
  - How is progress evaluated? (e.g., datasets, metrics)
  - Critical opinion & discussion (e.g., advantages disadvantages)
- Position your papers in respect to other papers
  - Commonalities
  - Differences
- Provide relevant background knowledge

Exemplary Workflow

1. Collect and prioritize

- Start with the given paper and establish basic concepts and terminology
- Identify further relevant papers through backward and forward citation and key-word search
- Prioritize by relevance and novelty; older papers for concepts newer for methods or results
- Categorize your findings

2. First-pass skim

- Read only title, abstract, introduction, conclusion, and figures/tables.
- Answer these quick questions in your notes:
  - What question is this paper trying to answer?
  - What's the proposed idea or technique?
  - What dataset or experiment validates it?
  - What are the key results (in one sentence)?
  - How is it related to your chosen subtopic?
- Record short bullet summaries
- After a >10 papers, you´ll notice recurring terms and debates (these will help to understand what to focus on)

3. Deep reading of core papers

- Select 6-8 papers most central to your subtopic and read them carefully
  - Read methods and experiments in detail
  - Re-express concepts in your own words
  - Compare results
  - Identify connections and differences between papers (e.g., method, architecture, training setup)
  - Identify limitations and open questions
- Deep reading means you can explain the idea on a whiteboard

4. Broader Contextualization

- Read secondary literature: survey papers, blog posts, technical reports to build intuition

5. Iterative Synthesis

- After 2-3 papers: Pause to integrate your findings into what you already have
  - Structure your current results into joint concepts/taxonomy
  - Write a short summary paragraph or update your notes

Resources

- **Literature search:**
  - https://www.semanticscholar.org
  - https://scholar.google.com/
- **Bibtex for the report:**
  - https://dblp.uni-trier.de/
- **High-res figures:**
  - On arxiv.org, under “Download: Other formats”,
- **Optional reading/ watching**
  - Presentation: MIT How to Speak (1h lecture)
  - Writing: Coursera Course - Writing in the Sciences (up to week 4, summary notes here)