



Short Advanced Studies (SAS)¹

Large Language Models – Architektur, Methoden und Praxis

Large Language Models verändern die Art, wie Unternehmen Wissen nutzen, Prozesse automatisieren und intelligente Anwendungen entwickeln. In diesem SAS erwerben Sie fundiertes Wissen zu Architektur, Methoden und Technologien moderner Sprachmodelle und setzen Chatbots, RAG-Systeme und KI-Agenten praxisnah um.

¹Short Advanced Studies (SAS) sind praxisnahe Kurzweiterbildungen, die sich an ein Fachpublikum richten, das sich im direkten Dialog mit Expert*innen neuen Herausforderungen stellen möchte (1–9 ECTS).

Inhaltsverzeichnis

1	Portrait	3
2	Berufsperspektiven	3
3	Zielpublikum	3
4	Ausbildungsziele	3
5	Voraussetzungen	3
6	Steckbrief	4
7	Inhalte und Lernziele	4
7.1	ML-Engineering und Technologien	4
7.2	Large Language Models verstehen und nutzen: Architektur, Methoden und Praxis	5
7	Kompetenznachweise	6
8	Dozierende	6
9	Organisation	6

Stand: 30.06.2026

1 Portrait

Large Language Models (LLMs) bilden die technologische Grundlage moderner KI-Anwendungen wie Chatbots, Copiloten und intelligente Assistenzsysteme. Dieses SAS vermittelt ein fundiertes Verständnis der zugrunde liegenden Konzepte, Architekturen und Methoden. Die Teilnehmenden lernen, LLM-basierte Anwendungen zu entwickeln, externe Wissensquellen über Retrieval Augmented Generation (RAG) einzubinden und Agentensysteme mit modernen Orchestrierungsframeworks umzusetzen.

2 Berufsperspektiven

Die Nachfrage nach Fachpersonen mit Kompetenzen im Bereich generativer KI wächst stark. Absolvent*innen dieses SAS können LLM-basierte Anwendungen konzipieren, evaluieren und in bestehende Systeme integrieren. Das erworbene Wissen unterstützt Sie in ihrer Aufgabe in den Bereichen Data Science, Software Engineering, AI Engineering.

3 Zielpublikum

- Softwareentwickler*innen, und IT-Fachleute, die LLM-basierte Anwendungen entwickeln möchten
- Data Engineers und Solution Architects
- Data Scientists und Machine-Learning-Praktiker*innen

4 Ausbildungsziele

Nach Abschluss des SAS:

- verstehen Sie die Architektur und Funktionsweise moderner Large Language Models.
- können Sie LLMs lokal oder cloudbasiert betreiben und in Anwendungen integrieren.
- entwickeln Sie RAG-basierte Chatbots mit Zugriff auf externe Wissensquellen.
- konzipieren und implementieren Sie einfache KI-Agenten unter Verwendung aktueller Frameworks und Standards wie MCP.
- können Sie Qualität, Leistungsfähigkeit und Einsatzgrenzen von LLM-Anwendungen bewerten.

5 Voraussetzungen

- Grundkenntnisse in Python oder einer vergleichbaren Programmiersprache.
- Basiswissen zu Machine Learning und neuronalen Netzen ist von Vorteil.
- Interesse an KI-Technologien und deren praktischer Anwendung.
- Bereitschaft zur Arbeit mit Entwicklungsumgebungen, APIs und Open-Source-Tools.

6 Steckbrief

Short Advanced Studies (SAS)	Large Language Models – Architektur, Methoden und Praxis
Titel/Abschluss	Short Advanced Studies (SAS) Large Language Models – Architektur, Methoden und Praxis
Anzahl ECTS	5 ECTS-Credits
Kosten	CHF 3'300
Unterrichtssprache	Deutsch / Literatur und Unterlagen teilweise in Englisch
Studienort	Aarbergstrasse 46, Biel / hybrid

7 Inhalte und Lernziele

7.1 ML-Engineering und Technologien

Sie lernen die wichtigsten Tools und Frameworks für maschinelles Lernen und KI kennen. Sie erfahren, wie Sie die richtigen Tools auswählen, diese in bestehende Systeme einbinden und Ihre KI-Modelle optimieren können. Zusätzlich behandeln wir kurz Entwicklungsumgebungen, Cloud-Dienste und skalierbare Infrastrukturen für Ihre KI-Projekte.

Die Themen und Technologien in diesem Kurs verändern sich schnell und werden für die Kursdurchführung zeitnah angepasst.

Lernziele	Die Teilnehmenden verstehen die grundlegenden Werkzeuge, Tools und die gängigen Frameworks für die Entwicklung und das Training neuronaler Netze. Sie können entscheiden, ob Modelle in der Cloud oder lokal gehostet werden sollen und kennen die relevanten Kriterien für diese Auswahl.
Themen und Inhalte	- Infrastruktur BFH nutzen - Technology Stack, Tools, - Google Colab, OLLAMA - Einblick in andere Entwicklungs-Frameworks - PyTorch, TensorFlow, Transformers - Python Environment aufbauen/aufsetzen - High-Performance-Computing: - o On-Premise - o Cloud - o Experimentell (bspw. Colab) - Huggingface als ModelHub - Evaluierung, Deployment (TensorBoard, TFX, Weights & Biases, MLFlow)
Lehrmittel	- Folien/Skript

7.2 Large Language Models verstehen und nutzen: Architektur, Methoden und Praxis

Grosse Sprachmodelle (Large Language Models LLMs) zielen darauf ab, komplexe textuelle Aufgaben mithilfe leistungsstarker neuronaler Netze zu bewältigen. Moderne LLMs wie GPT oder LLaMA haben den Bereich revolutioniert und bieten vielseitige Anwendungen. Durch Techniken wie Prompt Engineering, Feinabstimmung und semantische Suche können spezifische Fragestellungen rund um Textdaten effizient beantwortet werden.

Lernziele	Die Teilnehmenden - lernen die wichtigsten Konzepte moderner Sprachmodelle und deren praktische Anwendungen kennen. - Sie sind in der Lage, die Funktionsweise und Architektur von LLMs zu verstehen und deren Nutzen für verschiedene Aufgaben wie Textklassifikation, semantische Suche oder die Feinabstimmung auf spezifische Bedürfnisse einzuschätzen. - werden befähigt, Anwendungsfälle praktisch umzusetzen.
Themen und Inhalte	- Methoden - Foundation LLMs: Vortrainierte Sprachmodelle, die ein breites Spektrum an textbasierten Aufgaben lösen können. Sie können sowohl manuell durch menschliche Interaktion als auch programmatisch angesprochen werden (Beispiele: GPT, LLaMA). - Prompt Engineering: Die Kunst, gezielte Eingaben für Sprachmodelle zu entwickeln, um möglichst präzise und relevante Ausgaben zu erzielen (Techniken: Zero-shot, One-shot, Few-shot Learning, In-Context Learning). - Transformer-Architektur als Basis moderner LLMs. Das Verständnis ihrer Funktionsweise ermöglicht es, gezielt aufgabenspezifische Modelle zu bewerten und anzupassen (Encoder-Decoder, Encoder-only, Decoder-only; Konzepte: Embeddings). - Topic Modelling: Verfahren zur Gruppierung ähnlicher Dokumente oder Klassifikation und Clustering von Texten. - Semantische Suche und RAG (Retrieval-Augmented Generation): Ergänzt ein LLM um externes Wissen, beispielsweise um unternehmensspezifische zu Fragen zu beantworten, ohne dass das Modell explizit dafür trainiert wurde (Werkzeuge: LangChain). - Feinabstimmung (Finetuning): Weiterführendes Training eines Foundation-Modells auf spezifische Daten oder Aufgaben, um dessen Leistung zu optimieren (Methoden: Effizientes Lernen, Supervised Learning, Instruction Tuning, Reinforcement Learning mit menschlichem Feedback). - Herausforderungen und Grenzen von LLM: Qualität der Trainingsdaten, Bias, Halluzinationen, Schutz der Privatsphäre. - Anwendungen - Programmatische Anbindung von Sprachmodellen und deren taskbezogene Konfiguration - Optimierung von Aufgabenstellungen für LLMs - Textklassifikation - Anwendungen von RAG für spezifische Wissensabfragen - Feinabstimmung von Sprachmodellen für spezifische Aufgaben

Lehrmittel	- Folien/Skript - Unterstützende Lehrmittel: - Build an AI Agent (From Scratch); Hur, Song; Manning 2026; 978-1-633-43461-5 - Speech and Language Processing: D. Jurafsky and J. Martin; 2025; Website
------------	---

7 Kompetenznachweise

Bewertete Übungsaufgaben während des Kurses.

8 Dozierende

Vorname Name	Firma	E-Mail
Ilja Rasin	IBM / BFH	ilja.rasin@bfh.ch
Peter von Niederhäusern	BFH	peter.vonniederhaeusern@bfh.ch

9 Organisation

SAS-Leitung:

Arno Schmidhauser

Tel: [+41 31 848 32 75](tel:+41318483275)

E-Mail: arno.schmidhauser@bfh.ch

SAS-Administration:

Andrea Moser

Tel: [+41 31 848 32 11](tel:+41318483211)

E-Mail: andrea.moser@bfh.ch

Berner Fachhochschule

Technik und Informatik

Weiterbildung

Aarbergstrasse 46 (Switzerland Innovation Park Biel/Bienne)

2503 Biel/Bienne

Telefon +41 31 848 31 11

E-Mail: weiterbildung.ti@bfh.ch

bfh.ch/ti/weiterbildung

