Master's Thesis at the Pattern Recognition Lab Nr.:

**Development of a Local LLM Agent System**
**for Clinical Expert Support and Automation in MRI Planning for Radiation Therapy**
Entwicklung eines lokalen LLM-Agentensystems
zur klinischen Expertenunterstützung und Automatisierung der MRT-Planung in der Strahlentherapie

There is an unmet need to make healthcare workflows more efficient, safe, and personalized. Large Language Model (LLM) agents offer unprecedented capabilities for selectively collecting, processing, and combining patient-specific information from multiple sources to inform both clinical experts and automation solutions at the point of care. Recently, multiple capable open-weight reasoning LLMs have been released, which are optimally suited for local agentic use within a hospital environment [1–8]. This master thesis project aims to develop a local LLM agent system within the university hospital environment to inform and partially automate MRI treatment planning for radiation therapy. Being one of the most important treatment modalities for cancer, radiation therapy (RT) requires the combination of multifaceted information from diverse sources for optimal treatment planning. In this thesis, the LLM agent system will selectively retrieve and process patient information from the Picture Archiving and Communication System (PACS), Oncology Information System (OIS), and Radiology Information System (RIS), addressing two primary objectives. First, the system will deliver targeted, comprehensive and accurate patient data via an interactive LLM-powered dashboard to support clinical experts in making optimal treatment decisions. Second, it will facilitate automated treatment MRI planning by proposing appropriate MRI study protocols, examination decisions, and a personalized selection of MRI sequence protocols, thus enhancing both the efficiency and personalization of radiation therapy workflows.

The thesis will include the following points:

- Literature review on prior work, state-of-the-art open-weight LLMs, LLM agents, local LLM fine-tuning using hospital compute infrastructure and Retrieval-augmented generation (RAG) [1–10]. For this master thesis project, single and dual-48 Gb as well as dual-96 Gb GPU workstations are available.
- Set up test environment. Establishing read access to clinical databases (OIS, RIS) and PACS via python functions. Alternative: offline OIS / RIS (SQL)-database and PACS data access (using open-source PACS [ORTHANC or XNAT]). Fall-back option: local file explorer access to automatically downloaded patient data (DICOM, HTML, PDF, TXT format). Automatic downloading and script-based processing of past MRI scan data to generate ground truth reference, split into a validation and a test set. The LLM agent system will be developed based on validation data and tested on the separate test data set. Support will be provided for setting up the test environment and curating the validation and test data.
- LLM agent development using python, e.g. based on openai or langchain libraries [11, 12]. The local LLM will be served via ollama on a GPU workstation server (local gpt-oss 20b is already implemented within other LLM agent systems on-site as a reference) [13]. The LLM agent system should selectively retrieve and process available physician letters, prior imaging reports and reference image slice positions from PACS, RT indication, RT target volume, RT patient setup information, patient appointments and free-text notes from OIS as well as MRI scan information and patient warning messages from RIS via agent tools. Incorporation of institution-specific standards via integration of standard-operating-procedures (SOPs). Exploration of two underlying LLMs for the LLM agent system (e.g., Magistral 24B [multilingual reasoning with German texts] vs. gpt-oss 20B vs. Qwen3-30B-A3B)
- Development of the agent-fed dashboard application implemented using HTML and served over HTTP, that can be accessed at the point-of-care in the hospital (Prior HTML app is available on-site as a reference). The dash-board should depict the following information: Summary description of the patient case, fetched patient portrait photograph, patient oncologic diagnosis, oncologic stage, RT indication and target volume, summary of important side diagnoses for MRI (e.g. contrast media allergy), RT treatment setup description including fetched setup photograph, summary description of the anatomical location and extent of the tumor with fetched reference CT / MRI image slice from previous imaging studies. Overview of patient's next appointments.

Moreover, a proposal of the MRI exam protocol to be scanned, exam decisions and personalized sequence protocol selection as well as anatomic extent the sequence protocol field-of-views need to cover based on SOPs and extracted patient-information. A chat interface should be included for clinical experts to prompt the agent system to provide additional patient-specific information. Optional: Include feedback option to collect human feedback for later Reinforcement Learning (RL).

- Accuracy evaluation of the predicted MRI exam protocols, exam decisions and sequence protocol selection in a retrospective test set of 300 cases using already scanned MRI DICOM studies as ground truth reference. Accuracy evaluation of the additional output dashboard data with NLP metrics (ROUGE Score, BERT Score) and component-wise comparison of categorical and numerical output values using provided ground truth data.
- Point-of-care evaluation of the LLM-agent dashboard application and the LLM-recommended planning MRI exam settings by clinical experts, using a multidimensional rating scale.
- Statistical analysis and interpretation of the project results.

| | |
|---|---|
| *Academic advisors:* | PD Dr. med. Florian Putz, Dr.-Ing. Fabian Wagner, Annette Schwarz M.Sc., Prof. Dr.-Ing. habil. Andreas Maier |
| *Processor:* | |
| *Begin:* | 01.10.2025 |

## References


[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-T. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Advances in Neural Information Processing Systems (NeurIPS) 2020*, 2020.

[2] F. Putz, M. Haderlein, S. Lettmaier, S. Semrau, R. Fietkau, and Y. Huang, "Exploring the capabilities and limitations of large language models for radiation oncology decision support," *Int J Radiat Oncol Biol Phys*, vol. 118, no. 4, pp. 900–904, 2024. Epub 2024 Feb 22.

[3] Mistral AI, "Magistral-small-2506 model card." `https://huggingface.co/mistralai/Magistral-Small-2506`, June 2025. Hugging Face model card. Accessed 2025-08-24.

[4] OpenAI, "gpt-oss-120b & gpt-oss-20b model card." `https://openai.com/index/gpt-oss-model-card/`, Aug. 2025. OpenAI model card. Published August 5, 2025. Accessed 2025-08-24.

[5] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, and Z. Qiu, "Qwen3 technical report," 2025.

[6] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," 2023.

[7] M. Griot, J. Vanderdonckt, and D. Yuksel, "Implementation of large language models in electronic health records." Preprint, Research Square, 2025. Posted July 4, 2025. DOI: 10.21203/rs.3.rs-7029913/v1. CC BY 4.0.

[8] A. Wada, Y. Tanaka, M. Nishizawa, A. Yamamoto, T. Akashi, A. Hagiwara, Y. Hayakawa, J. Kikuta, K. Shimoji, K. Sano, K. Kamagata, A. Nakanishi, and S. Aoki, "Retrieval-augmented generation elevates local llm quality in radiology contrast media consultation," *NPJ Digital Medicine*, vol. 8, p. 395, 2025. Published July 2, 2025. Open Access (CC BY 4.0).

[9] Y. Hou, C. Bert, A. Gomaa, G. Lahmer, D. Höfler, T. Weissmann, R. Voigt, P. Schubert, C. Schmitter, A. Depardon, S. Semrau, A. Maier, R. Fietkau, Y. Huang, and F. Putz, "Fine-tuning a local llama-3 large language model for automated privacy-preserving physician letter generation in radiation oncology," *Frontiers in Artificial Intelligence*, vol. 7, p. 1493716, Jan 2025. eCollection 2024.

[10] A. Palepu, V. Dhillon, P. Niravath, W.-H. Weng, P. Prasad, K. Saab, R. Tanno, Y. Cheng, H. Mai, E. Burns, Z. Ajmal, K. Kulkarni, P. Mansfield, D. Webster, J. Barral, J. Gottweis, M. Schaekermann, S. S. Mahdavi, V. Natarajan, A. Karthikesalingam, and T. Tu, "Exploring large language models for specialist-level oncology care," 2024.

[11] H. Chase *et al.*, "Langchain: A framework for developing applications with large language models." `https://github.com/hwchase17/langchain`, 2025. Accessed: 2025-08-24; replace with release tag or DOI if available.

[12] OpenAI, "openai-python: The official python library for the openai api." `https://github.com/openai/openai-python`, 2025. Accessed: 2025-08-24; include the specific release tag or commit SHA if desired.

[13] Ollama, "Ollama: Run and manage language models locally." `https://github.com/ollama/ollama`, 2025. Accessed: 2025-08-24; no official academic paper found as of access date.