



CCA – COMPETENCE CENTRE

**HTL Anichstraße**



# Diplomarbeit

## Tapyre

Entwicklung eines KI-integrierten Produktivitätstools mit Plugin-System

Eingereicht von

**Christian Vorhofer**  
**Raphael Ladinig**

Eingereicht bei

**Höhere Technische Bundeslehr- und Versuchsanstalt**  
**Anichstraße**

Abteilung für Wirtschaftsingenieure/Betriebsinformatik

Betreuer

GREINÖCKER Albert, Mag. Dr. DI

Innsbruck, April 2026

---

Abgabevermerk:

Betreuer/in:

Datum:



## Kurzfassung / Abstract

Eine Kurzfassung ist in deutscher sowie ein Abstract in englischer Sprache mit je maximal einer A4-Seite zu erstellen. Die Beschreibung sollte wesentliche Aspekte des Projektes in technischer Hinsicht beschreiben. Die Zielgruppe der Kurzbeschreibung sind auch Nicht-Techniker! Viele Leser lesen oft nur diese Seite.

### Beispiel für ein Abstract (DE und EN)

Die vorliegende Diplomarbeit beschäftigt sich mit verschiedenen Fragen des Lernens Erwachsener – mit dem Ziel, Lernkulturen zu beschreiben, die die Umsetzung des Konzeptes des Lebensbegleitenden Lernens (LBL) unterstützen. Die Lernfähigkeit Erwachsener und die unterschiedlichen Motive, die Erwachsene zum Lernen veranlassen, bilden den Ausgangspunkt dieser Arbeit. Die anschließende Auseinandersetzung mit Selbstgesteuertem Lernen, sowie den daraus resultierenden neuen Rollenzuschreibungen und Aufgaben, die sich bei dieser Form des Lernens für Lernende, Lehrende und Institutionen der Erwachsenenbildung ergeben, soll eine erste Möglichkeit aufzeigen, die zur Umsetzung dieses Konzeptes des LBL beiträgt. Darüber hinaus wird im Zusammenhang mit selbstgesteuerten Lernprozessen Erwachsener die Rolle der Informations- und Kommunikationstechnologien im Rahmen des LBL näher erläutert, denn die Eröffnung neuer Wege zur orts- und zeitunabhängiger Kommunikation und Kooperation der Lernenden untereinander sowie zwischen Lernenden und Lernberatern gewinnt immer mehr an Bedeutung. Abschließend wird das Thema der Sichtbarmachung, Bewertung und Anerkennung des informellen und nicht-formalen Lernens aufgegriffen und deren Beitrag zum LBL erörtert. Diese Arbeit soll

einerseits einen Beitrag zur besseren Verbreitung der verschiedenen Lernkulturen leisten und andererseits einen Reflexionsprozess bei Erwachsenen, die sich lebensbegleitend weiterbilden, in Gang setzen und sie somit dabei unterstützen, eine für sie geeignete Lernkultur zu finden.

This thesis deals with the various questions concerning learning for adults – with the aim to describe learning cultures which support the concept of live-long learning (LLL). The learning ability of adults and the various motives which lead to adults learning are the starting point of this thesis. The following analysis on self-directed learning as well as the resulting new attribution of roles and tasks which arise for learners, trainers and institutions in adult education, shall demonstrate first possibilities to contribute to the implementation of the concept of LLL. In addition, the role of information and communication technologies in the framework of LLL will be closer described in context of self-directed learning processes of adults as the opening of new forms of communication and co-operation independent of location and time between learners as well as between learners and tutors gains more importance. Finally the topic of visualisation, validation and recognition of informal and non-formal learning and their contribution to LLL is discussed.

Gliederung des Abstract in **Thema, Ausgangspunkt, Kurzbeschreibung, Zielsetzung**.

**Projektergebnis** Allgemeine Beschreibung, was vom Projektziel umgesetzt wurde, in einigen kurzen Sätzen. Optional Hinweise auf Erweiterungen. Gut machen sich in diesem Kapitel auch Bilder vom Gerät (HW) bzw. Screenshots (SW). Liste aller im Pflichtenheft aufgeführten Anforderungen, die nur teilweise oder gar nicht umgesetzt wurden (mit Begründungen).

# Erklärung der Eigenständigkeit der Arbeit

## EIDESSTATTLICHE ERKLÄRUNG

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche erkenntlich gemacht habe. Meine Arbeit darf öffentlich zugänglich gemacht werden, wenn kein Sperrvermerk vorliegt.

Ort, Datum

Verfasser 1

Ort, Datum

Verfasser 1



# Inhaltsverzeichnis

<b>Abstract</b>	<b>ii</b>
<b>1 Einführung in Neuronale Netzwerke</b>	<b>1</b>
1.1 Künstliche Neuronen . . . . .	1
1.2 Feed-Forward Neural Networks (FNN) . . . . .	2
1.3 Rekurrente Neuronale Netze (RNN, LSTM) . . . . .	3
1.4 Die Transformer-Architektur . . . . .	4
1.5 Bedeutung von Transformern für LLMs und Embeddings . .	5
<b>2 Einführung in Natural Language Processing (NLP)</b>	<b>7</b>
2.1 Klassische NLP-Ansätze . . . . .	7
2.2 Einführung in Embeddings . . . . .	8
2.3 Word Embeddings: Word2Vec und GloVe . . . . .	9
2.4 Kontextualisierte Embeddings . . . . .	9
2.5 Embeddings mit der Transformer-Architektur . . . . .	10
2.6 Relevanz für Tapyre Paper Search . . . . .	10
<b>3 Einführung in Agentic AI</b>	<b>11</b>
3.1 ReAct: Reasoning + Acting . . . . .	11
3.2 Tool Usage . . . . .	12
3.3 Model Context Protocol (MCP) . . . . .	12
3.4 Agent-to-Agent Kommunikation . . . . .	13
3.5 RAG: Retrieval-Augmented Generation . . . . .	13
3.6 Multi-Agent Systems . . . . .	14
<b>4 Grundkonzepte der verwendeten Technologien</b>	<b>15</b>
4.1 Docker und Containerisierung . . . . .	15
4.2 MySQL als relationale Datenbank . . . . .	16

4.3	Qdrant und Approximate Nearest Neighbor Search . . . . .	17
4.3.1	Speicherstruktur . . . . .	17
4.3.2	Approximate Nearest Neighbor (ANN) . . . . .	17
4.3.3	Ähnlichkeitsmaße . . . . .	18
4.4	Flask und REST-APIs . . . . .	18
4.5	PyTorch und GPU-Beschleunigung . . . . .	19
4.6	Zusammenfassung . . . . .	20
<b>5</b>	<b>Entwicklung von Tapyre als Agentic-AI-System</b>	<b>21</b>
5.1	Abstraktion der LLM-Schnittstelle . . . . .	22
5.2	Der Agent und der ReAct-Loop . . . . .	23
5.3	Plugins als Tools: Lose Kopplung durch Interfaces . . . . .	24
5.4	Dynamisches Laden der Plugins . . . . .	25
5.5	Beispiel: AppPlugin zur Steuerung lokaler Anwendungen . . . . .	26
5.6	Zusammenspiel von Agent, Plugins und ReAct-Loop . . . . .	28
<b>6</b>	<b>Architektur und Implementierung von Tapyre Paper Search</b>	<b>29</b>
6.1	Abstraktion der Datenquellen . . . . .	30
6.2	ArXiv als konkrete Datenquelle . . . . .	31
6.3	PDF-Verarbeitung und Textextraktion . . . . .	34
6.4	Abstraktion der Embedding-Erzeugung . . . . .	37
6.5	Specter2 als semantisches Embedding-Modell . . . . .	37
6.6	Abstraktion der Datenhaltung . . . . .	39
6.7	MySQL für strukturierte Metadaten . . . . .	40
6.8	Qdrant als Vektordatenbank . . . . .	43
6.9	Pipeline zur Orchestrierung des Gesamtprozesses . . . . .	45
6.10	Zusammenspiel der Komponenten . . . . .	48
6.11	Analyse der Performance-Charakteristika . . . . .	48
6.11.1	Limitierungen bei der semantischen Suche . . . . .	48
6.11.2	Anfängliche Implementierungsineffizienzen und Optimierungen . . . . .	49
6.11.3	Periodische Verarbeitungseinbrüche durch geplante Unterbrechungen . . . . .	50
6.11.4	Kumulative Verarbeitung und Gesamtstabilität des Systems . . . . .	50
6.11.5	Zusammenfassende Einordnung . . . . .	51



**Literaturverzeichnis**

**61**

Christian Vorhofer  
Raphael Ladinig



# Appendix



# Tabellenverzeichnis



# Abbildungsverzeichnis

6.1	Anzahl der pro Tag verarbeiteten wissenschaftlichen Publikationen im Testbetrieb . . . . .	49
6.2	Kumulative Anzahl der verarbeiteten wissenschaftlichen Publikationen . . . . .	51





# Listings

5.1	Abstrakte LLM-Schnittstelle . . . . .	22
5.2	OllamaLLM als konkrete Implementierung . . . . .	22
5.3	Abstrakte Agent-Schnittstelle . . . . .	23
5.4	PluginAgent mit ReAct-Agententyp . . . . .	23
5.5	Abstrakte Plugin-Basisklasse . . . . .	24
5.6	Dynamischer PluginLoader . . . . .	25
5.7	AppPlugin als konkretes Plugin . . . . .	26
6.1	Abstrakte Schnittstelle für Datenquellen . . . . .	30
6.2	arXivDataProvider zur Anbindung der arXiv-API . . . . .	31
6.3	PDF-zu-Text-Konvertierung mit PyMuPDF . . . . .	34
6.4	Abstrakte Embedder-Schnittstelle . . . . .	37
6.5	Specter2Embedder zur Erzeugung semantischer Vektoren . . . . .	38
6.6	Abstrakte Datenbankschnittstelle . . . . .	39
6.7	MySQL-Datenbankanbindung für Paper-Metadaten . . . . .	40
6.8	Qdrant-Datenbank für semantische Suche . . . . .	43
6.9	Pipeline zur Verarbeitung und Indexierung von Papers . . . . .	46



# Literaturverzeichnis

- Ashish Vaswani, Noam Shazeer, N. P. J. U. L. J. A. N. G. L. K. I. P. (2017), 'Arxiv', <https://arxiv.org/abs/1706.03762>. Zugriff 2025.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M. et al. (2020), Language models are few-shot learners, in 'Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)'.
- Chen, W. (2023), 'Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks', *arXiv preprint arXiv:2305.10200* .
- Cohan, A., Feldman, S., Beltagy, I., Downey, D. & Weld, D. (2020), 'Specter: Document-level representation learning for scientific papers', *arXiv preprint arXiv:2004.07180* .  
**URL:** <https://arxiv.org/abs/2004.07180>
- Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018), 'Bert: Pre-training of deep bidirectional transformers for language understanding', *arXiv preprint arXiv:1810.04805* .  
**URL:** <https://arxiv.org/abs/1810.04805>
- Docker Inc. (2025), 'Docker engine security', <https://docs.docker.com/engine/security/>. Accessed 2025.
- Du, N., Liu, H., Li, Y. et al. (2023), 'Improving factuality and reasoning in language models through multiagent debate', *arXiv preprint arXiv:2305.14325* .
- Fielding, R. T. (2000), Architectural Styles and the Design of Network-based Software Architectures, PhD thesis, University of California, Irvine.
- Gamma, E., Helm, R., Johnson, R. & Vlissides, J. (1994), *Design Patterns: Elements of Reusable Object-Oriented Software*, Addison-Wesley.

- Goodfellow, I., Bengio, Y. & Courville, A. (2016), *Deep Learning*, MIT Press.  
<http://www.deeplearningbook.org>.
- Gray, J. & Reuter, A. (1992), *Transaction Processing: Concepts and Techniques*, Morgan Kaufmann.
- Hochreiter, S. & Schmidhuber, J. (1997), *long short-term memory*. <https://www.bioinf.jku.at/publications/older/2604.pdf>.
- Hong, B., Tang, Y., Li, H. et al. (2023), 'Metagpt: Meta programming for multi-agent collaborative framework', *arXiv preprint arXiv:2308.00352*.
- Jurafsky, D. & Martin, J. H. (2023), *Speech and Language Processing*. Online version.  
**URL:** <https://web.stanford.edu/jurafsky/slp3/>
- LangChain (2023), 'Langchain documentation', <https://python.langchain.com/>. Accessed 2025.
- Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. (1998), 'Gradient-based learning applied to document recognition', *Proceedings of the IEEE* **86**(11), 2278–2324.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N. et al. (2020), 'Retrieval-augmented generation for knowledge-intensive nlp tasks', *arXiv preprint arXiv:2005.11401*.  
**URL:** <https://arxiv.org/abs/2005.11401>
- Malkov, Y. A. & Yashunin, D. A. (2020), 'Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **42**(4), 824–836.
- Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013), 'Efficient estimation of word representations in vector space', *arXiv preprint arXiv:1301.3781*.  
**URL:** <https://arxiv.org/abs/1301.3781>
- Nandakumar, K., Cohan, A., Feldman, S., Downey, D. & Beltagy, I. (2023), Specter2: Building better document-level representations, in 'Findings of the Association for Computational Linguistics (ACL)'.  
**URL:** <https://arxiv.org/abs/2305.12248>
- NVIDIA Corporation (2025), *CUDA C++ Programming Guide*. Zugriff am: 05.12.2025.

- OpenAI (2024), 'Model context protocol', <https://github.com/modelcontextprotocol>. Accessed 2025.
- Oracle Corporation (2024), *MySQL 8.4 Reference Manual: InnoDB Index Types*. Accessed 2025.  
**URL:** <https://dev.mysql.com/doc/refman/8.4/en/innodb-index-types.html>
- Paszke, A., Gross, S., Massa, F. et al. (2019), Pytorch: An imperative style, high-performance deep learning library, in 'Advances in Neural Information Processing Systems 32 (NeurIPS 2019)'.
- Percona (2024), 'Understanding mysql indexes: Types, benefits, and best practices'. Accessed 2025.  
**URL:** <https://www.percona.com/blog/understanding-mysql-indexes-types-best-practices/>
- Qdrant Technologies (2024a), 'Hnsw indexing fundamentals', <https://qdrant.tech/course/essentials/day-2/what-is-hnsw/>. Accessed 2025.
- Qdrant Technologies (2024b), 'Qdrant concepts: Collections', <https://qdrant.tech/documentation/concepts/collections/>. Accessed 2025.
- Qdrant Technologies (2024c), 'Qdrant concepts: Indexing', <https://qdrant.tech/documentation/concepts/indexing/>. Accessed 2025.
- Qdrant Technologies (2024d), 'Qdrant documentation', <https://qdrant.tech/documentation/>. Accessed 2025.
- Qdrant Technologies (2024e), 'Similarity search in qdrant', <https://qdrant.tech/documentation/concepts/search/>. Accessed 2025.
- Quirós, G. (2024), 'How containers work: Layers, overlayfs, namespaces & cgroups'. Accessed 2025.  
**URL:** <https://k8studio.io/tutorials/container-architecture-namespaces-cgroups-overlayfs/>
- Ronacher, A. & Contributors, F. (2024), 'Flask documentation', <https://flask.palletsprojects.com/>. Zugriff am: 05.12.2025.
- Schick, T., Dwivedi-Yu, J., Vu, T. et al. (2023), 'Toolformer: Language models can teach themselves to use tools', *arXiv preprint arXiv:2302.04761*.  
**URL:** <https://arxiv.org/abs/2302.04761>

- Touvron, H., Lavril, T., Izacard, G., Martinet, X. et al. (2023), 'Llama: Open and efficient foundation language models', *arXiv preprint arXiv:2302.13971* .
- Wang, Z., Zhu, C., Lin, Y. & Zhou, J. (2024), 'A survey on agentic large language models', *arXiv preprint arXiv:2401.05561* .
- Yao, S., Deng, J. Z., Zhou, J., Wang, A., Lo, Y. & Goodman, N. (2022), 'React: Synergizing reasoning and acting in language models', *arXiv preprint arXiv:2210.03629* .  
**URL:** <https://arxiv.org/abs/2210.03629>