BHT Berliner Hochschule für Technik

Wissensmanagement in der IT-Administration durch wissensbasierte Large Language Models

vorgelegt von

Friederike Buchner

EDV.Nr.:xxxxxx

dem Fachbereich VI – Informatik und Medien
der Berliner Hochschule für Technik Berlin
vorgelegte Arbeit
zur Absolvierung des Moduls

Wissenschaftliches Seminar

im Studiengang

Medieninformatik (Online)

Tag der Abgabe 3. November 2025

Studiere Zukunft

Version 0.1α
letzte Änderung: 3. November 2025

Gutachter
Prof. Dr. S. Edlich Berliner Hochschule für Technik

Kurzfassung

Die Kurzfassung gibt ein kurzes und prägnantes Bild der gesamten Arbeit. Sie soll den Leser neugierig machen und klarmachen, was zu erwarten ist. Erreichte Ergebnisse werden kurz umrissen.

Abstract

Bachelor and Master-Thesises usually are often written in german. Nevertheless, their content may be intersting for people being not able to read german. In order to awaken their interest in this topic, an abstract in english is given. The experimental results and analysis are shown in short

Inhaltsverzeichnis

Kapitel 1

Einleitung

Nutzung von KI (GPT-5):

- Übersetzung (Englisch-Deutsch):
  - self-supervision
  - downstream tasks
  - homogenization
  - emergence
  - transfer learning

Kapitel 2

Theoretische und technische Grundlagen

2.1 Wissensmanagement in Organisationen

2.2 Large Language Models (LLMs) und Foundation Models

Foundation Models nach [Bommasani et al.] sind Modelle, die üblicherweise selbstüberwacht (*self-supervised*) auf umfassenden Daten trainiert sind und auf eine große Anzahl an nachgelagerten Aufgaben (*downstream tasks*) angepasst werden können. Aktuelle Beispiele beinhalten BERT, GPT-3 und CLIP. Von einem technologischen Standpunkt her sind *Foundation Models* nicht neu, da sie auf tiefen neuronalen Netzwerken und selbstüberwachtem Lernen basieren, was beides bereits seit Jahrzehnten existiert. Beachtenswert sind *Foundation Models* heutzutage deshalb, weil sich ihre schiere Größe in den letzten Jahren vervielfacht hat und sie somit alle Vorstellungen dessen, was man vor wenigen Jahren für möglich hielt, sprengen. GPT-3 beispielsweise hat 175 Milliarden Parameter und kann durch natürlichsprachige Prompts so angepasst werden, dass es eine passable Leistung in vielfältigen Aufgaben zeigt, obwohl es nicht explizit für diese Aufgaben trainiert wurde [Bommasani et al. , S. 3].
Nach technischen Gesichtspunkten funktionieren *Foundation Models* durch Transferlernen (*transfer learning*) und Skalierung (*scale*). Transferlernen bedeutet, das „Wissen“, was in einer Anwendung (bspw. Bilderkennung) erlernt wurde, auf eine andere Aufgabe (bspw. das Erkennen von Aktivitäten in Videos) zu übertragen. Innerhalb des *Deep Learning* ist das Vortraining (*pretraining*) der vorherrschende Ansatz für Transferlernen: ein Model trainiert eine Ersatzaufgabe und wird dann via *fine-tuning* für die eigentlich relevante nachgelagerte Aufgabe angepasst. Zusammen mit der Skalierung von *Foundation Models* entsteht nun eine sehr mächtige Komination. Hierfür werden drei entscheidende Punkte wichtig: die Verbesserungen in Computer Hardware, die Entwicklung der Transformer Architektur und die Verfügbarkeit von viel mehr Trainingsdaten [Bommasani et al. , S. 4].
Letzteres kann nicht deutlich genug hervorgehoben werden: die Wichtigkeit des Vorhandenseins von Daten und die Fähigkeit, sich diese zunutze zu machen. Transferlernen durch annotierte Datensätze war jahrelang die gängige Praxis. Die hohen Kosten von Annotationen, insbesondere von hochqualitativen, händisch erzeugten Annotationen, haben jedoch eine natürliche Grenze in der Skalierung von Trainingsdaten dargestellt. Im selbstüberwachten Lernen ergibt sich die Ersatzaufgabe für das Vortraining automatisch aus unannotierten Daten. Selbstüberwachte Aufgaben sind nicht nur besser skalierbar und ausschließlich abhängig von unannotierten Daten, sondern sie zwingen das Modell dazu, Teile der Eingabe vorherzusehen, was sie reichhaltiger und potentiell nützlicher machen als Modelle, die in einem begrenzteren Sprachaum trainiert sind [Bommasani et al. , S. 4].

Selbstüberwachtes Lernen war zunächst eine Unterdisziplin von NLP, die sich parallel zu anderen Entwicklungen ergab. Ab der Einführung des BERT-Modells [Devlin et al.] 2019 wurde selbstüberwachtes Lernen zur Norm in NLP. Die Akzeptanz, dass ein einzelnes Modell derart nützlich über eine weite Bandbreite von Aufgaben sein könnte, markiert den Beginn der Ära von *Foundation Models* [Bommasani et al. , S. 5].
Homogenisierung ist ein Ergebnis der Konsolidierung von Systemen für Maschinelles Lernen über eine weite Palette an Anwendungen. Es ermöglicht das Erledigen vieler Aufgaben aber bildet auch *single points of failure* [Bommasani et al. , S. 3]. *Foundation Models* haben ein nie zuvor gesehenes Maß an Homogenisierung herbeigeführt, da fast alle *state-of-the-art* NLP-Modelle aus einem von wenigen Modellen wie BERT, GPT o.ä. hervorgehen. Dadurch können alle NLP-Anwendungen direkt von Verbesserungen in *Foundation Models* profitieren. Es birgt aber auch die Gefahr, dass alle KI-Systeme dieselben problematischen Verzerrungen (*biases*) einiger weniger *Foundation Modeld* erben.
Ein zweites Charakteristikum von *Foundation Models* ist die Emergenz. Das bedeutet, dass das Verhalten eines Systems implizit induziert ist, anstatt explizit konstruiert. Das zeigt sich, indem ein System (zur Überraschung seiner Schaffer*innen) Verhaltensweisen oder Fähigkeiten aufweist, die nicht von außen definiert wurden, sondern die sich eher als Nebenprodukt zum hauptsächlichen Einsatzzweck ergeben. Es ist gleichzeitig die Quelle wissenschaftlicher Erregung sowie Besorgnis über unerwartete Konsequenzen [Bommasani et al. , S. 3]. Emergenz wird umso bedeutender, je größer das Modell skaliert ist. Während GPT-2 mit 1,5 Milliarden Parametern trainiert wurde, wurde GPT-3 mit 175 Milliarden Parametern trainiert, was kontextsensitives Lernen ermöglichte, in welchem das Sprachmodell einfach auf eine nachgelagerte Aufgabe angepasst wird, indem es mit einer natürlichsprachlichen Beschreibung einer Aufgabe (*prompt*) versorgt wird. Dies war eine emergente Fähigkeit, die weder speziell trainiert noch überhaupt antizipiert wurde [Bommasani et al. , S. 5].
Homogenisierung und Emergenz interagieren miteinander auf potenziell beunruhigende Art und Weise. Homogenisierung kann potenziell enorme Vortaile für viele Domänen bringen, in denen aufgabenspezifische Daten knapp sind. Auf der anderen Seite kann jeder Fehler im Modell blind von allen angepassten Modellen geerbt werden. Da die Macht von *Foundation Models* viel mehr in ihren emergenten Qualitäten als in ihrer expliziten Konstruktion steckt, sind die existierenden Modelle schwer zu verstehen und haben unvorhergesehene Fehlverhalten. Da Emergenz substanzielle Unsicherheiten über die Fähigkeiten und Fehler von *Foundation Models* erzeugt, ist mit der umfassenden Homogenisierung über diese Modelle hinweg ein erhebliches Risiko verbunden. Dieses Risiko zu mitigieren ist eine der zentralen Herausforderungen in der weiteren Entwicklung von *Foundation Models* aus einer ethischen sowie aus einer KI-Sicherheitsperspektive [Bommasani et al. , S. 6].
Zur Bezeichnung von *Foundation Models* und zur Abgrenzung von Sprachmodellen allgemein kann man sagen, dass der Geltungsbereich von *Foundation Models* schlichtweg weit über die Grenzen von Sprache hinaus reicht. Es wurden auch andere Bezeichnungen wie beispielsweise *General-Purpose Model* oder *Multi-Purpose Model* in Betracht gezogen, da sie den Aspekt, dass diese Modelle vielfältige nachgelagerte Aufgaben bewältigen können, besser einfängen. Sie täuschen aber darüber hinweg, dass *Foundation Models* unfertig sind und weiter angepasst werden müssen. Weitere Namensvorschläge wie *Task-agnostic Model* würden zwar die Art des Trainings wiederspiegeln, aber nicht die Relevanz für weitere nachgelagerte Aufgaben. Es wurde sich also für den Begriff *Foundation* (engl. für Basis, Grundlage) entschieden, da ein *Foundation Model* an sich unfertig ist, aber als allgemeine Grundlage dient, aus der vielfältige aufgabenspezifische Modelle durch Anpassung entstehen können. Gleichzeitig weist der Begriff *Foundation* auch auf die Wichtigkeit von architektonischer Stabilität, funktionaler Sicherheit (engl. *safety*) sowie dem Schutz vor Angriffen (engl. *security*) hin. So wie schlecht konstruierte Fundamente fast schon eine Garantie für strukturelles Versagen sind, sind gut ausgeführte Fundamente verlässliche Grundlagen für zukünftige Anwendungen. Zu betonen ist weiterhin, dass momentan die Natur oder

Qualität dieser Art von Fundamenten, die *Foundation Models* bieten, nicht in Gänze verstanden wird und dass nicht einwandfrei beurteilt werden kann, ob diese Fundamente verlässlich sind, oder nicht.

2.3 Retrieval-Augmented Generation (RAG)

LLMs haben neben ihrem bemerkenswerten Erfolg auch signifikante Grenzen, speziell in domänenspezifischen oder wissensintensiven Aufgaben. Eins der größten Probleme ist das „Halluzinieren“ [Zhang et al.] beim Verarbeiten von Anfragen, die Informationen betreffen, die nicht in den Trainingsdaten enthalten waren. Um diese Herausforderungen zu bewältigen, werden LLMs per *Retrieval-Augmented-Generation* (RAG) erweitert, indem relevante Inhalte mithilfe semantischer Ähnlichkeitsberechnungen aus externen Wissensbasen abgerufen werden. Indem externes Wissen referenziert wird, reduziert RAG effektiv das Problem, faktisch inkorrekte inhalte zu generieren. Die Integration in LLMs ist mittlerweile weit verbreitet, was RAG zu einer Schlüsseltechnologie im Voranbringen von Chatbots und der Eignung von LLMs für Anwendungen in der realen Welt gemacht hat [Gao et al.].
Die Erforschung von RAG traf mit der Entwicklung der Transformer Architektur zeitlich aufeinander. Zu Beginn lag der Fokus darauf, Sprachmodelle durch zusätzliche Wissensquellen zu verbessern, insbesondere durch die Integration externer Informationen in vortrainierte Modelle (*Pretrained Models*, PTMs). Mit dem Aufkommen von ChatGPT gab es einen Wendepunkt: Große Sprachmodelle (LLMs) zeigten nun ihre Fähigkeit zum *In-Context Learning* (ICL), also dazu, zur Laufzeit neues Wissen aus Eingabekontexten aufzunehmen und zu verwenden. Das führte die RAG-Forschung dahin, bessere Informationen für LLMs bereitzustellen, um komplexere und wissensintensive Aufgaben in der Inferenz-Phase (also während der Antwortgenerierung) beantworten zu können. Mit voranschreitender Forschung war RAG dann nicht mehr auf die Inferenz-Phase beschränkt, sondern fügte sich immer mehr in LLM *fine-tuning*-Techniken, also das gezielte Nachtrainieren der Modelle mit domänenspezifischen oder aufgabenspezifischen Daten, ein [Gao et al.].

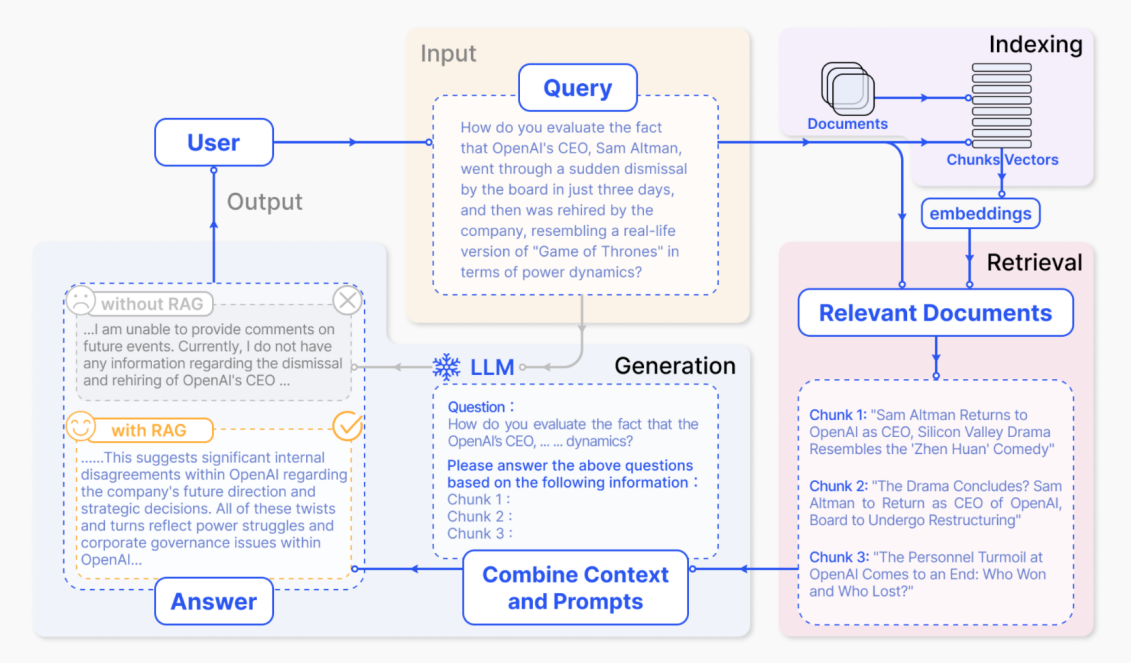

Abbildung 2.1: Überblick über die Funktionsweise von RAG nach [Gao et al. , S. 3]

In Abbildung 2.1 ist die grundsätzliche Funktionsweise von RAG, angewendet auf die Aufgabe der Fragenbeantwortung, dargestellt. Sie besteht aus drei Schritten:

1. *Indexing*: Dokumente werden in Abschnitte (engl. *chunks*) unterteilt, in Vektoren kodiert und in einer Vektordatenbank gespeichert
2. *Retrieval*: die relevantesten Top k Abschnitte werden abgerufen, basierend auf semantischer Ähnlichkeit
3. *Generation*: die ursprüngliche Frage wird gemeinsam mit den abgerufenen Abschnitten an ein LLM übergeben, um eine Antwort zu generieren

Weiterhin: Naive RAG (s.o.) + Schwächen (Retrieval Challenges, Generation Difficulties, Augmentation Hurdles)
Advances RAG (Pre-Retrieval Process, Post-Retrieval Process)
Modular RAG (New Modules, New Patterns)
RAG vs. Fine-Tuning

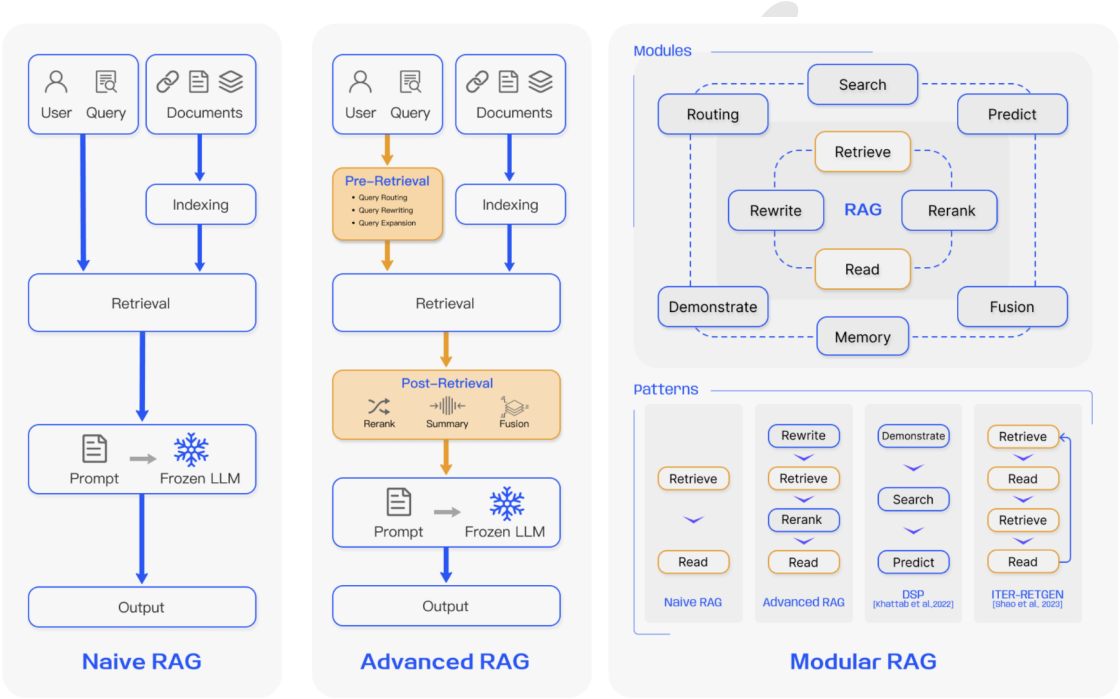

Abbildung 2.2: Überblick über *Naive RAG*, *Advanced RAG* und *Modular RAG* [Gao et al. , S. 3]

2.4 Praktische Referenzen und Systeme

Kapitel 3

Konzeption des RAG-Prototyps

3.1 Zieldefinition und Anforderungen

3.2 Architekturentwurf

Kapitel 4

Implementierung eines Minimalbeispiels

4.1 Technologiewahl und Setup

4.2 Umsetzung der Komponenten

Kapitel 5

Evaluation

5.1 Evaluationsdesign

5.2 Ergebnisse (Dummy-Data)

Kapitel 6

Diskussion

Kapitel 7

Fazit und Ausblick

Literaturverzeichnis

[Bommasani et al.] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Arx, S. v., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., und Liang, P., On the opportunities and risks of foundation models.

[Devlin et al.] Devlin, J., Chang, M.-W., Lee, K., und Toutanova, K., BERT: Pre-training of deep bidirectional transformers for language understanding.

[Gao et al.] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., und Wang, H., Retrieval-augmented generation for large language models: A survey.

[Zhang et al.] Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Xu, C., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., und Shi, S., Siren's song in the AI ocean: A survey on hallucination in large language models.