KomMKonLLM

Endbericht | Call 19 | Projekt ID 7409

Inhalt

1 Einleitung

- Wie konsistent ist ein bestimmtes LLM, wenn Teile von Prompts durch Synonyme ersetzt werden?
- Welches LLM ist für eine bestimmte Fragengruppe am besten geeignet?
- Welche Fähigkeiten hat ein LLM in Bezug auf Anweisungen, die sein Ausgabeformat einschränken?

Mit der zunehmenden Integration von Large Language Models (LLMs) in Produktentwicklung und Dienstleistungen wird deren Konsistenz und Verlässlichkeit ein zentraler Erfolgsfaktor. Gerade in unternehmenskritischen Anwendungen müssen LLMs zuverlässig auf ähnliche Eingaben konsistente Antworten liefern, auch wenn sich die exakte Formulierung ändert.

Das Projekt **KomMKonLLM** widmet sich dieser Herausforderung, indem es systematische Methoden für das Konsistenztesten von LLMs betrachtet. In **KomMKonLLM** wenden wir kombinatorische Testmethoden aus dem Softwaretesten an, um Konsistenztests für LLMs zu generieren.

Das Projektziel ist der Aufbau einer automatisierten Plattform, die mithilfe kombinatorischer Testmethoden LLMs auf Konsistenz testet. Zu den Kernfunktionen zählen die Erstellung, Durchführung und Auswertung von Tests sowie die Bereitstellung von Ergebnissen über eine benutzerfreundliche Schnittstelle. Zielgruppen sind Wissenschaft, Forschung, LLM-Nutzer:innen sowie Unternehmen, die alle LLMs sicher und effizient einsetzen möchten.

2 Projektbeschreibung

2.1 Projektziel

Das Projektziel war eine einfach zu verwendende Lösung für das systematische Testen von Large Language Models (LLMs) hinsichtlich ihrer semantischen Konsistenz zur Verfügung zu stellen.

2.2 Was bedeutet Konsistenztesten?

Das Konsistenztesten von LLMs adressiert das Problem, sicherzustellen, dass LLMs zuverlässig auf verschiedene semantisch äquivalente Eingaben reagieren. Da LLMs oft komplexe und undurchsichtige Strukturen haben, können sie inkonsistente oder unerwartete Antworten auf ähnliche Eingaben geben. Diese Inkonsistenzen erschweren es, LLMs in Anwendungen einzusetzen, wo Verlässlichkeit entscheidend ist. Die Herausforderung liegt darin, geeignete Testmethoden zu entwickeln, die systematisch die Konsistenz der Modelle bewerten.

Die besondere Herausforderung beim Testen der semantischen Konsistenz von LLMs ist es die vielen möglichen semantisch äquivalenten Variationen von Formulierungen von

Eingabeabfragen abzudecken. Um diesem Problem zu begegnen, haben wir automatisierte kombinatorische Testmethoden eingesetzt. Diese Methoden ermöglichen es, den modellierten Eingaberaum – bestehend aus unterschiedlichen semantisch äquivalenten Formulierungen – gezielt und strukturiert abzudecken. Dadurch lassen sich systematische Inkonsistenzen in den LLM-Antworten identifizieren, die bei rein manueller (oder zufälliger) Testfallauswahl leicht übersehen werden könnten. Kombinatorisches Testen erlaubt es also, trotz begrenzter Ressourcen eine möglichst breite (vorgegebene) Abdeckung der semantischen Varianten zu erreichen.

2.3 Zielgruppen

Ganz allgemein zählen natürlich alle Nutzer:innen von LLMs zur Zielgruppe der in unserem Projekt entwickelten Technologie. Im Detail ist unsere Technologie an die folgenden drei Zielgruppen adressiert:

- *Zielgruppe 1: Entwickler:innen und Open Source Community*

  Die auf den ersten Blick überraschenden Fähigkeiten von LLMs haben diese Technologie nicht nur in das Zentrum des öffentlichen Interesses gebracht, sondern auch zu einem regelrechten LLM-Hype bei Softwareentwickler:innen geführt. Diese stehen bei der Realisierung ihrer Softwareprojekte vor der Entscheidung, welche Kandidaten-LLMs sie in ihre Open Source-Lösungen einbetten sollen. Da diese Entscheidung die Qualität des Softwareprojekts (un)mittelbar beeinflusst ist es wichtig, dass das LLM konsistent korrekte Antworten liefert. Dies wird umso mehr verdeutlicht, wenn das Softwareprojekt der Entwickler:innen von technisch wenig erfahrene Nutzer:innen verwendet wird, welche Ausgaben unreflektiert übernehmen.
  Unsere Software-Lösung mit der integrierten, anpassungsfähigen Analyse samt grafischer Aufbereitung ermöglicht es Entwickler:innen eine fundiertere Entscheidung bei der Wahl von LLMs zu treffen.

- *Zielgruppe 2: LLM-Nutzer:innen und Unternehmen*

  Zahlreiche Unternehmen sind dabei, LLMs in ihre Prozesse zu integrieren, Prozesse zu automatisieren, zu verbessern oder zu beschleunigen. Dies trifft sowohl auf firmeninterne Prozesse, etwa Textverarbeitung in der internen Kommunikation, als auch externe Prozesse, wie etwa die Kundenkommunikation zu. Das Bedürfnis der Organisationen ist, dass Nutzer:innen des durch LLM gestützten Prozesses akkurate und *konsistente Antworten* erhalten, unabhängig von der genauen Formulierung des Prompts. Ähnliches gilt für private LLM-Nutzer:innen.
  Unsere Software-Lösung mit der integrierten, anpassungsfähigen Analyse samt grafischer Aufbereitung ermöglicht es Unternehmen eine fundiertere Wahl zu treffen, wenn es um die Integration von LLMs in ihre Produktentwicklung oder geht.

- *Zielgruppe 3: Forscher:innen*

Die in der jüngsten Zeit signifikante Verbesserung von LLMs haben diese Technologie auch in den Mittelpunkt der Aufmerksamkeit vieler Forscher:innen gebracht, was sich weltweit in einer Fülle an Forschungsarbeiten zu diesem Thema bemerkbar macht. Gegenstand deren Forschung sind dabei neben der Verbesserung von LLMs auch das Entwickeln von Methoden für das Testen von LLMs, insbesondere das *Konsistenztesten von LLMs.* Entwickelte Testverfahren dienen nicht zuletzt dazu die Qualität von LLMs zu verbessern sowie deren Limitierungen besser zu verstehen. *Unsere Server-software gemeinsam mit dem Testdatensatz kann für Forscher:innen gleichermaßen einen Referenzpunkt, wie einen Startpunkt für Weiterentwicklungen darstellen.*

2.4 Inhaltlicher Überblick über die Projektergebnisse

Die **KomMKonLLM** Server-Software (https://github.com/KomMKonLLM/KomMKonLLM) stellt eine serverseitige Softwarelösung zur Durchführung kombinatorischer Konsistenztests von LLMs bereit. Die Architektur des Systems ist modular aufgebaut: Verschiedene Komponenten übernehmen spezialisierte Aufgaben und kommunizieren über standardisierte HTTP-Schnittstellen miteinander. Ein Überblick über den gesamten Test-Prozess wird in *„Abbildung 2 Ein Überblick über den Gesamtprozess.“* gegeben. Zur Orchestrierung dieser Komponenten wird Docker-Compose eingesetzt, was eine einfache Installation und Skalierbarkeit ermöglicht. Die zentrale Kontrolllogik zur Generierung, Ausführung und Auswertung der Testfälle ist in Python implementiert. Die Testergebnisse werden strukturiert in einer relationalen Datenbank abgelegt, was eine nachvollziehbare und wiederholbare Analyse erlaubt. Für Nutzer:innen steht zudem ein JupyterLab-Notebook als interaktive Oberfläche zur Verfügung, das eine einfache Nutzung sowie flexible Erweiterbarkeit durch eigenen Code ermöglicht. Die verwendete Systemarchitektur (siehe „Abbildung 1 Veranschaulichung der Gesamtarchitektur des **KomMKonLLM-Frameworks**.“) wird im folgenden Kapitel im Detail erklärt.

Für den als *Open Data* bestehend aus kombinatorisch generierten Testfällen zur Konsistenzevaluierung von LLMs zur Verfügung gestellten Testdatensatz wurde ein entsprechender Beispieldatensatz von Fragen samt (1) gekennzeichneten richtigen Antworten, sowie die im **KomMKonLLM** Testprozess (2) verwendeten Synonymen, (3) den verwendeten Covering Arrays und (4) den LLM-Ausgaben für 6 LLMs sowie (5) deren Boolesche Interpretation erzeugt. Der Datensatz ist online verfügbar unter: https://zenodo.org/records/15209547.

3 Verlauf der Arbeitspakete

3.1 Arbeitspaket 1 - Detailplanung und Formales am Projektstart

Das **erste Arbeitspaket** wurde im Dezember 2024 erfolgreich abgeschlossen. Insbesondere sind folgende Ergebnisse entstanden:

- Beidseitig unterschriebener Fördervertrag;

- Detailprojektplan (inklusive Plan der einzelnen Arbeitspakete) erstellt und abgenommen;
- Liste der Projektergebnisse mit zugehöriger Lizenz erstellt;
- Projektwebseite (https://www.netidee.at/kommkonllm) von **KomMKonLLM** innerhalb von Netidee in Betrieb genommen und ersten Blogeintrag (https://www.netidee.at/kommkonllm/ueberblick) erstellt;
- Beantragung von erster Förderrate und deren Genehmigung nach Feedback.

Es wurden in AP1 einige Planstunden nicht verbraucht; diese wurden auf AP4 verschoben, wo sich ein leicht erhöhter Implementierungsaufwand abgezeichnet hatte, welcher dann auch eingetreten ist.

3.2 Arbeitspaket 2 - Technische Planung

Das **zweite Arbeitspaket** beschäftigte sich mit einer Anforderungsanalyse von **KomMKonLLM**, der technischen Gesamtarchitektur und der Auswahl technischer Komponenten (wie – unter anderem – Frameworks, Programmiersprachen und Schnittstellen) und wurde Anfang Jänner 2025 abgeschlossen[1]. Es wurden in AP2 einige Planstunden nicht verbraucht; diese wurden auf AP4 verschoben, wo sich ein leicht erhöhter Implementierungsaufwand abgezeichnet hatte, welcher dann auch eingetreten ist.

Konkret wurden in AP2 die folgenden Aufgaben abgearbeitet und die dazugehörigen Zwischenziele (bzw. Zielgruppen) erreicht:

- Analyse der Anforderungen für ein LLM-Testframework
- Entwicklung einer reproduzierbaren technischen Gesamtarchitektur
    - Zielgruppe: Projektmitarbeiter, Softwareentwickler:innen
- Auswahl geeigneter Technologien, Frameworks und Programmiersprachen

Die erhaltenen Erkenntnisse aus AP2 haben wir in unserem zweiten Blogpost (https://www.netidee.at/kommkonllm/architektur-technologien) zugänglich gemacht, den wir im Folgenden kurz zusammenfassen. Einen Überblick über die technische Architektur zeigt das

[1] Aufgrund von Urlauben and Krankenständen über Weihnachten und Silvester 2024/2025 endete AP2 statt zu Weihnachten 2024 erst Anfang 2025.

folgende Diagramm:

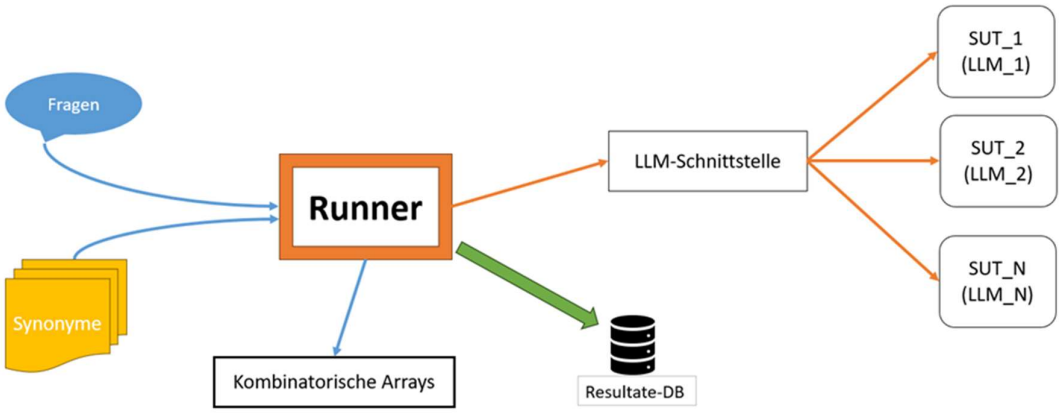

*Abbildung 1 Veranschaulichung der Gesamtarchitektur des **KomMKonLLM**-Frameworks.*

Eine detaillierte Beschreibung der Aufgaben bzw. Anforderungen an diese Komponenten ist online (https://github.com/KomMKonLLM/KomMKonLLM/blob/main/docs/Architecture.pdf) zu finden.

Weiters wählten wir zu verwendende Technologien in **KomMKonLLM** unter Berücksichtigung folgender Kriterien aus:

- **Popularität bei Entwicklern**: Um die Verbreitung und Weiterentwicklung von **KomMKonLLM** zu vereinfachen, sollten Technologien weit genutzt werden.
- **Aktive Entwicklung**: Um das Risiko von Software-Obsoleszenz von verwendeten Komponenten zu minimieren, sollten eingebundene Technologien nicht nur zum jetzigen Zeitpunkt unter aktiver Entwicklung sein, sondern es sollte auch eine aktive Weiterentwicklung in der Zukunft anzunehmen sein. Für diese Entscheidungen wurden auch die Einschätzungen von CrOSSD[2] herangezogen.
- **Lizenzkompatibilität**: Ausgewählte Technologien dürfen keine Einschränkungen auf die Lizenzierung von **KomMKonLLM**[3] haben.

Die anhand dieser Kriterien ausgewählten Technologien und Bibliotheken sind online (https://github.com/KomMKonLLM/KomMKonLLM/blob/main/docs/Selected%20Technologies.pdf) zusammengefasst.

3.3 Arbeitspaket 3 - Methodische Entwicklung LLM Testing Framework

[2] https://crossd.tech/ (Netidee gefördertes Projekt 2022: https://www.netidee.at/crossd; bzw. Nachfolge-Projekt **CrOSSD2 | netidee** Netidee Call #19 (2024): **https://www.netidee.at/crossd2**).
[3] Die Lizenz von Softwareergebnissen von **KomMKonLLM** ist *MIT license*.

Das dritte Arbeitspaket wurde (nach kurzer Verlängerung) Anfang Februar 2025 fertiggestellt; wobei einige Planstunden nicht verbraucht wurden, diese wurden auf AP4 verschoben, wo sich ein leicht erhöhter Implementierungsaufwand abgezeichnet hatte, welcher dann auch eingetreten ist. Das AP3 beschäftigte sich mit der Methodik zur Erzeugung kombinatorischer Konsistenztests für LLMs. Wir haben unseren gewählten methodischen Ansatz von theoretischer Seite in zwei Blog-Posts beschrieben (Teil 1: https://www.netidee.at/kommkonllm/methodik-von-kommkonllm-teil-1-von-2; und Teil 2: https://www.netidee.at/kommkonllm/methodik-von-kommkonllm-teil-2-von-2) und in einem Weiteren (https://www.netidee.at/kommkonllm/beispiel-fuer-das-erzeugen-kombinatorischer-konsistenzfragen) ein konkretes Beispiel durchgeführt. Ein weiterer Blogpost (https://www.netidee.at/kommkonllm/kombinatorisches-testen-aller-kuerze) gibt einen Überblick über kombinatorisches Testen.

Konkret wurden in AP3 die folgenden Aufgaben abgearbeitet und die dazugehörigen Zwischenziele (bzw. Zielgruppen) erreicht:

- Entwicklung kombinatorischer Eingabemodelle auf Basis von Frage-Antwort-Korpora und Synonymdatenbanken
    - Zielgruppe: Projektmitarbeiter sowie (Weiter-) Entwickler:innen
- Modellierung, Generierung und Normalisierung von Testfällen
    - Zielgruppe: Projektmitarbeiter sowie (Weiter-) Entwickler:innen
- Entwicklung eines semantischen Test-Orakels zur Bewertung von LLM-Antworten
    - Zielgruppe: Projektmitarbeiter sowie (Weiter-) Entwickler:innen

Nachfolgend geben wir eine Zusammenfassung von AP3. Die in **KomMKonLLM** implementierte Methode zur Erzeugung von Konsistenztests für LLMs kann auf jede binäre Entscheidungsfrage (vereinfacht als Ja/Nein- beziehungsweise richtig/falsch-Frage beschreibbar) in natürlicher Sprache angewendet werden. Wir verwenden Wortersetzungen durch Synonyme und kombinatorische Methoden, um strukturiert und in (kombinatorisch-) messbarer Diversität abgewandelte Fragen zur Konsistenzevaluierung zu erstellen. Ausgehend von einer gegebenen binären Entscheidungsfrage, gliedert sich der Gesamtprozess zur Erzeugung von kombinatorischen Konsistenztests in folgende Schritte:

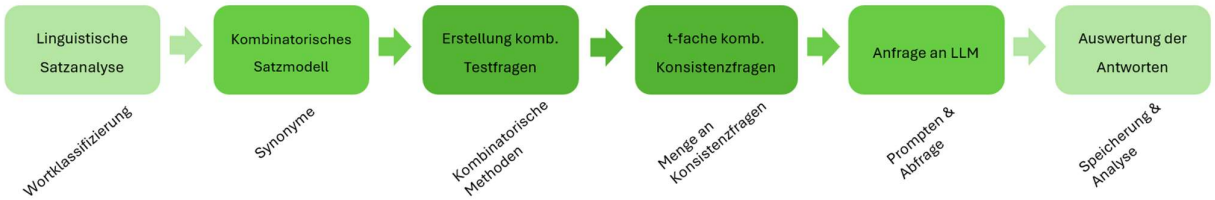

Abbildung 2 Ein Überblick über den Gesamtprozess.

- **„Linguistische Satzanalyse**“: In diesem Schritt wenden wir NLP-Techniken an, um aus der gegebenen binären Entscheidungsfrage ein diskretisiertes Satzmodell zu erzeugen.
- „**Kombinatorisches Satzmodell**“: In diesem Schritt verwenden wir Synonym-Datenbanken, um für die in der Frage vorkommenden Wörter der ausgewählten lexikalischen Klassen eine bestimmte Anzahl an Synonymen auszuwählen, wobei jedes Wort auch selbst als erstes Synonym in die entsprechende Liste hinzugefügt wird.
- „**Erstellung kombinatorischer Testfragen**“: Basierend auf dem abgeleiteten IPM[4] verwenden wir nun *abdeckende Array-Strukturen* (engl.: covering arrays), welche in minimalisierter Zeilenanzahl das Auftreten von allen t-fachen Parameter-Wertekombinationen garantieren.
- **„t-fache kombinatorische Konsistenzfragen“**: Jede Zeile der im vorherigen Schritt erzeugten Array-Struktur kann nun wieder in eine Frage zurückgeführt werden, indem die Wörter durch ihre in der Zeile angegebenen Synonyme ersetzt werden.
- „**Anfrage an LLM**“: Jede der so erzeugten Fragen wird nun mit einem speziellen Prompt, der eine binäre Antwort des LLMs forcieren soll, an ein LLM gesendet und das Frage-Antwort Paar in einer Datenbank gespeichert.
- **„Auswertung der Antworten**“: Alle Frage/Resultat-Paare, welche basierend auf der abdeckenden Array-Struktur erstellt wurden, können nun ausgewertet werden.

3.4 Arbeitspaket 4 - Implementierung LLM Testing Framework

Das vierte Arbeitspaket wurde geändert schon mit Ende Dezember 2024 gestartet, um vorab die Kompatibilität möglicher (Code-) Komponenten mit dem entwickelten kombinatorischen Testverfahren zu klären. Dies ermöglichte eine frühe Koordination zwischen dem Arbeitspaket 3 und Arbeitspaket 4. Der betriebene Mehraufwand für Arbeitspaket 4 konnte allerdings durch Zeiteinsparungen in den anderen Arbeitspaketen ausgeglichen werden und AP4 wurde im März 2025 fertiggestellt. Insbesondere ergab sich daraus eine vom Plan abweichende Aufteilung an Arbeitsstunden auf die Arbeitspakete und die Projektmitarbeiter.

AP4 umfasste die praktische Umsetzung des Testing-Frameworks. Es wurden zentrale Komponenten wie ein Prozess-Orchestrator, ein Modellgenerator sowie Schnittstellen zu LLMs implementiert. Das entwickelte System ist in der Lage, Testfälle in ausführbare Prompts zu übersetzen, diese an LLMs zu übergeben und die Rückgaben automatisiert auf Konsistenz zu prüfen und zu speichern.

[4] Die Bezeichnung *IPM* steht für *„input parameter model“* und ergibt sich aus dem vorherigen Schritt.

Konkret wurden in AP4 die folgenden Aufgaben abgearbeitet und die dazugehörigen Zwischenziele (bzw. Zielgruppen) erreicht:

- Umsetzung von Komponenten wie Prozess-Orchestrator, Modellgenerator und Testfallübersetzer
    - Zielgruppe: Projektmitarbeiter sowie (Weiter-) Entwickler:innen
- Anbindung verschiedener LLMs und deren Einbettung in den Testprozess
    - Zielgruppe: Projektmitarbeiter sowie (Weiter-) Entwickler:innen
- Speicherung und automatisierte Evaluierung der Testergebnisse
    - Zielgruppe: Projektmitarbeiter sowie (Weiter-) Entwickler:innen und Nutzer:innen von LLMs, die verschiedene LLMs nach deren Konsistenz vergleichen wollen.

Es folgt ein Überblick über den Implementierungvelauf des LLM Test Frameworks **KomMKonLLM**:

- Der zentrale Runner, sowie die Datenbank-Anbindung für das Abspeichern der Resultate waren zügig fertiggestellt.
- Die Implementierung der Schnittstelle zur Testfallgenerierung fiel bereits umfangreicher aus als geplant, da wir nicht nur eine Schnittstelle zu einem CA-Generator erstellt haben, sondern drei Schnittstellen zu frei verfügbaren Generatoren realisiert haben.
- Die diversen LLMs wurden über Ollama (https://ollama.com/), einem Framework das verschiedene open-source Large Language Models lokal verfügbar und über eine Kommandozeile oder API nutzbar macht, integriert. Die Verwendung von Ollama hat das Setup der LLMs vereinfacht, und die Schnittstellen zu den LLMs vereinheitlicht. Letztlich hat sich allerdings die Implementierung der Schnittstelle zu Ollama als aufwendiger herausgestellt als ursprünglich gedacht und geplant war. Darin begründet sich auch der knapp 20-prozentige Überschuss (circa eineinhalb Vollzeit-Äquivalent-Wochen) an Arbeitszeit für Arbeitspaket 4.

3.5 Arbeitspaket 5 - Technische Dokumentation und Release-Vorbereitung

Im fünften Arbeitspaket erfolgte die technische Dokumentation aller Systembestandteile von **KomMKonLLM**. Es wurden umfassende Architektur- und Setup-Dokumentationen erstellt, inklusive Readme-Dateien für Entwickler und Anwender. Die Codebasis und Dokumentationen wurden für eine Open-Source-Veröffentlichung vorbereitet.

In AP5 wurden die folgenden Tätigkeiten durchgeführt und Dokumente erstellt:

- Ausführliche Dokumentation des Systems für Entwickler:innen und Anwender:innen
    - Zielgruppe: Software Entwickler:innen und Nutzer:innen von LLMs, d.h. sowohl Firmen als auch Privatpersonen
- Bereitstellung von Setup- und Nutzungsanleitungen
    - Zielgruppe: Software Entwickler:innen und Nutzer:innen von LLMs
- Veröffentlichung der entwickelten Software und Dokumentation als Open Source
    - Zielgruppe: Software Entwickler:innen und Nutzer:innen von LLMs
- Veröffentlichung des erstellten kombinatorischen Testdatensatzes
    - Zielgruppe: LLM Nutzer:innen und Entwickler:innen

Die technische Dokumentation, die Entwickler:innen Dokumentation, sowie die Nutzer:innen Dokumentation ist plangemäß verlaufen. Die Ergebnisse sind unter https://github.com/KomMKonLLM/KomMKonLLM frei zugänglich, insbesondere eine technische Dokumentation für Nutzer:innen und Entwickler:innen (https://github.com/KomMKonLLM/KomMKonLLM/blob/main/README.md), sowie eine technische Dokumentation explizit für Entwickler:innen zum Weiterentwickeln (https://github.com/KomMKonLLM/KomMKonLLM/blob/main/HACKING.md). Außerdem haben wir einen umfangreichen User Guide[5] erstellt, welcher eine kompakte Zusammenfassung bietet, die Hintergründe des kombinatorischen Testens anhand eines motivierenden Beispiels erläutert und eine ausführliche Nutzer:innen-Anleitung mit Überblick über die Systemarchitektur sowie ein ausgearbeitetes Beispiel zur graphischen Auswertung enthält. Um LLM Nutzer:innen eine einfache Möglichkeit für das Konsistenztesten von LLMs zu eröffnen wurden kombinatorisch generierte Testfälle online zur Verfügung gestellt[6].

Die Zuweisung der Planstunden auf die Projektmitarbeiter ist in der Durchführung aus praktischen Gründen vom Plan abgewichen, insofern, dass einige Stunden umverteilt wurden und das insgesamt leicht weniger Stunden als ursprünglich geplant in AP5 verbraucht wurden. Weiters wurde AP5 schon Mitte Februar 2025 gestartet und dann erfolgreich Ende März 2025 abgeschlossen.

3.6 Arbeitspaket 6 - Dokumentation und Formales am Projektende

Abschließend wurde im Rahmen des letzten Arbeitspakets der Projektendbericht verfasst und gemeinsam mit der Endabrechnung an netidee übermittelt. Die finalen Projektergebnisse wurden öffentlich zugänglich gemacht (Open Source) und die Projektwebsite aktualisiert.

[5] https://www.netidee.at/sites/default/files/2025-05/KomMKonLLM_UserGuide.pdf
[6] https://zenodo.org/records/15209547

In AP6 wurden die folgenden Dokumente erstellt:

- Zusammenfassung der Projektergebnisse und Endbericht
    - Zielgruppe: netidee Administration, Internetstiftung Administration, Projektmitarbeiter, Nutzer:innen, Entwickler:innen, Forscher:innen
- Öffentliche Zugänglichmachung aller relevanten Materialien
    - Zielgruppe: Nutzer:innen, Entwickler:innen, Forscher:innen
- Kommunikation des Projekterfolgs über Blogbeiträge und Vorträge
    - Zielgruppe: Nutzer:innen, Entwickler:innen, Forscher:innen

Aufgrund von Krankheit, Urlaub sowie Feiertagen hat sich die inhaltliche Fertigstellung des Endberichts auf Ende Mai verschoben. Danach haben die administrativen Vorgaben zur Projektabrechnung mehr Zeit als erwartet in Anspruch genommen, da einige Missverständnisse und Fehlkommunikation bezüglich der Erstellung der geforderten Lohnabrechnung geklärt werden mussten, wodurch sich die Abgabe des Endberichts auf Ende Juni 2025 verzögert hat. Diese zeitliche Verschiebung hat sich jedoch nicht negativ auf den Inhalt bzw. die Qualität des Projekts ausgewirkt.

Die Zuweisung der Planstunden auf die Projektmitarbeiter ist in der Durchführung aus praktischen Gründen vom Plan abgewichen, insofern, dass einige Stunden umverteilt wurden und insgesamt etwas mehr Stunden als ursprünglich geplant in AP6 verbraucht wurden.

4 Umsetzung Förderauflagen

Im unterzeichneten Fördervertrag wurden *keine* Förderauflagen festgelegt.

5 Liste Projektendergebnisse

1	Projektzwischenbericht	CC BY 4.0	https://www.netidee.at/sites/default/files/2025-02/prj7409_Call19_Zwischenbericht_V01.pdf
2	Projektendbericht	CC BY 4.0	https://www.netidee.at/kommkonllm
3	Entwickler_innen-DOKUMENTATION	CC BY 4.0	https://www.netidee.at/sites/default/files/2025-05/HACKING_on_KomMKonLLM.pdf

4	Anwender_innen-DOKUMENTATION	CC BY 4.0	https://www.netidee.at/kommkonllm
5	Veröffentlichungsfähiger Einseiter / Zusammenfassung	CC BY 4.0	*https://www.netidee.at/sites/default/files/2025-05/KomMKonLLM_Einseiter.pdf*
6	Dokumentation Externkommunikation zur Erreichung Sichtbarkeit /Nachhaltigkeit	CC BY 4.0	*https://www.netidee.at/kommkonllm* Die Dokumentation Externkommunikation zur Erreichung Sichtbarkeit ist als Teil des Endberichts, in Kapitel **7 Öffentlichkeitsarbeit/Vernetzung,** zu finden**.**
7	Server-Lösung für kombinatorische Konsistenztests von LLMs	MIT	*https://github.com/KomMKonLLM/KomMKonLLM*
8	Repository von kombinatorischen Testfällen zur Konsistenzevaluierung von LLMs	MIT	*https://zenodo.org/records/15209547*

6 Verwertung der Projektergebnisse in der Praxis

Die Verwendung von unzuverlässigen Sprachmodellen kann schnell zu finanziellen Schäden und dem Verlust von Reputation führen, wenn automatisch generierte Antworten falsche oder rechtlich problematische Informationen als Fakten präsentieren. Das entwickelte **KomMKonLLM**-Framework hilft Entwickler:innen und Unternehmen, vor und während der Implementierung von LLM-basierten Produkten und Anwendungen existierende Modelle auf ihre Konsistenz zu testen und so eine informierte Auswahl zu treffen. Besonders wichtig dabei ist, dass einerseits anwendungsspezifische Frage-Antwort-Paare für die Evaluierung benutzt werden können und andererseits durch die Generierung von kombinatorischen Testsätzen ein garantiertes Maß an Diversität von Formulierungen erreicht wird.

Für Endanwender ist die Verwendung unserer Implementierung hauptsächlich sinnvoll, um für eigene Projekte ein zuverlässiges Sprachmodell zu finden und problematische Formulierungen bzw. etwaige Probleme mit spezifischen Modellen früh zu erkennen.

Eine beispielhafte Evaluierung von sechs LLMs anhand von 27 ursprünglichen Fragen (die anhand unseres Ansatzes auf knapp 550 Testeingaben diversifiziert wurden) ist unter

https://zenodo.org/records/15209547 verfügbar und wurde mit Stand Mitte Mai 2025 schon über 500 mal heruntergeladen. Pro Sprachmodell wurden dabei die klassischen Kennzahlen Precision, Recall und F1-Score ermittelt. Eine Erweiterung um zusätzliche Kennzahlen ist durch die interaktive Entwicklung in einem JupyterLab-Notebook, welches in einem normalen Browser verändert und ausgeführt werden kann, mit Einsteiger-Programmierkenntnissen möglich. Zudem wird in der Dokumentation für Entwickler:innen erklärt, wie Anbindungen an weitere LLMs oder Test-Generatoren implementiert werden können.

Nachdem unsere Open Source-Veröffentlichung erst einige Wochen zurückliegt und keinerlei Registrierung notwendig ist, um die **KomMKonLLM** Software zu verwenden, sind uns noch keine externen Anwender bekannt. Wir freuen uns jedoch auf Feedback und eventuell sogar Weiterentwicklungen seitens der Community.

7 Öffentlichkeitsarbeit/ Vernetzung

Die **KomMKonLLM** Projekt-Homepage (https://www.netidee.at/kommkonllm) auf der Netidee-Homepage ist die zentrale Anlaufstelle für Informationen über **KomMKonLLM**, wobei es aber auch eine Kurzbeschreibung (https://www.sba-research.org/research/projects/kommkonllm/) von **KomMKonLLM** auf der Homepage des Forschungszentrums SBA Research[7] sowie einen Eintrag in der Projektliste der MATRIS-Forschungsgruppe gibt (https://matris.sba-research.org/current-projects/).

Zum Zeitpunkt des Endberichts haben wir sieben Blog-Beiträge innerhalb der Netidee-**KomMKonLLM** Homepage erstellt:

- Überblick (https://www.netidee.at/kommkonllm/ueberblick);
- Architektur & Technologien (https://www.netidee.at/kommkonllm/architektur-technologien);
- Methodik von KomMKonLLM: Teil 1 von 2 (https://www.netidee.at/kommkonllm/methodik-von-kommkonllm-teil-1-von-2);
- Methodik von KomMKonLLM: Teil 2 von 2 (https://www.netidee.at/kommkonllm/methodik-von-kommkonllm-teil-2-von-2);
- Beispiel für das Erzeugen von kombinatorischen Konsistenzfragen (https://www.netidee.at/kommkonllm/beispiel-fuer-das-erzeugen-kombinatorischer-konsistenzfragen);

[7] Das COMET-Zentrum SBA Research (SBA-K1 NGC) wird im Rahmen von COMET – Competence Centers for Excellent Technologies durch BMIMI, BMWET und das Land Wien gefördert. COMET wird durch die FFG abgewickelt.

- Kombinatorisches Testen in aller Kürze (https://www.netidee.at/kommkonllm/kombinatorisches-testen-aller-kuerze);
- Konsistenztesten von LLMs: Die praktische Implementierung (https://www.netidee.at/kommkonllm/konsistenztesten-von-llms-die-praktische-implementierung).

Zusätzlich zu den netidee-Blogposts, wurden im Rahmen des Projekts folgende Maßnahmen zur Öffentlichkeitsarbeit und Vernetzung gesetzt:

- Ein offizieller Bluesky-Account (https://bsky.app/profile/kommkonllm.bsky.social) wurde eingerichtet, um über Projektfortschritte, Erkenntnisse und Veranstaltungen zu informieren. Wir planen, diesen Account auch nach Abschluss des netidee Projekts weiterzuführen.
- Im Jänner 2025 erfolgte eine Online-Präsentation des Projektes **KomMKonLLM** an das Management und die wissenschaftliche Leitung des Forschungszentrums SBA Research.
- Am 25. Februar 2025 wurde im Rahmen der Veranstaltung „*SBA Security Meetup hosted by Dynatrace!*“ (https://www.meetup.com/login/?returnUri=https%3A%2F%2Fwww.meetup.com%2Fsecurity-meetup-by-sba-research%2Fevents%2F305866708%2F) ein (externer) Vortrag (https://www.sba-research.org/2025/03/26/sba-security-meetup-dynatrace/) abgehalten, um das Projekt einem sicherheitsinteressierten Fachpublikum, Entwickler:innen und LLM Nutzer:innen, vorzustellen, zur Diskussion anzuregen und Weiterentwicklungsvorschläge einzuholen. Bei diesem Vortrag gab es gut 60 Zuhörer:innen.
- Ein (interner) Vortrag über **KomMKonLLM** wurde SBA Research-weit im März 2025 abgehalten, mit inhaltichem speziell gelegtem Fokus für Forscher:innen und LLM Nutzer:innen. Bei diesem Vortrag gab es 76 Zuhörer:innen.

Weitere Maßnahmen zur Vernetzung mit der Forschungs- sowie Open-Source-Communities sind geplant, insbesondere im Rahmen von Workshops und Beiträgen auf Fachkonferenzen. Dazu zählt auch der Netidee Spring Talk Call 19 (abgehalten am 3. Juni 2025) sowie Science- & Tech-Outreach Aktivitäten analog zu Security Meetups.

Da die Anzahl der Follower unseres eigens angelegten Social-Media Accounts noch gering ist, haben wir vor, Informationen über den erfolgreichen Abschluss von **KomMKonLLM**, sowie Verweise auf die netidee-Projektwebsite und des Github-Repository zusätzlich zu den offiziellen und professionellen Kanäle auch über unsere privaten Social-Media Accounts zu verbreiten. Außerdem haben wir vor, LLM-Nutzer:innen und Software-Entwickler:innen bei

Fachveranstaltungen zu erreichen (siehe dazu auch „9 Geplante Aktivitäten nach netidee-Projektende).

Essenziell sind natürlich auch die Github-Website (https://github.com/KomMKonLLM/) von **KomMKonLLM** samt Github-Repository (https://github.com/KomMKonLLM/KomMKonLLM), sowie das Repository von kombinatorischen Testfällen zur Konsistenzevaluierung von LLMs (https://zenodo.org/records/15209547).

Lessons learned

Bei unseren Outreach-Aktivitäten haben wir die Beobachtung gemacht, dass “analoge” soziale Netzwerke besser funktionieren als “digitale” soziale Netzwerke – zumindest bisher und zumindest was Rückmeldungen und Ideen bezüglich des **KomMKonLLM**-Frameworks betrifft. Bei unseren Vorträgen, ob bei uns am Forschungszentrum, oder beim Security Meet-Up bei Dynatrace, haben wir zahlreiche Wortmeldungen mit Fragen, Ideen und Anregungen zur Funktionsweise und zur Weiterentwicklung des **KomMKonLLM**-Frameworks erhalten.

Auf der anderen Seite sehen wir die Vorteile der Skalierbarkeit durch digitale Netzwerke. So wurden unsere Testdatensätze auf Zenodo zum Zeitpunkt Mitte Mai bereits über 500-mal heruntergeladen.

Einige der oben erwähnten, persönlich eingeholten Anregungen, sowie unsere eigenen Ideen zur Weiterentwicklung möchten wir in Zukunft aufgreifen, um das **KomMKonLLM**-Framework zu verbessern, beziehungsweise um seine Einsatzmöglichkeiten zu erweitern.

Weitere Aktivitäten der *Öffentlichkeitsarbeit* befinden sich derzeit in der Planung, siehe dazu **9 Geplante Aktivitäten nach netidee-Projektende**.

8 Eigene Projektwebsite

Neben der netidee-Projektseite wird aktuell *keine* weitere Projektseite für **KomMKonLLM** betrieben.

9 Geplante Aktivitäten nach netidee-Projektende

Wir planen unsere Outreach-Aktivitäten weiterzuführen, um das **KomMKonLLM**-Framework zu bewerben und zu demonstrieren, insbesondere da es nun einsatzfähig ist. Dabei fokussieren wir uns auf folgende Zielgruppen:

1.) **LLM-Entwickler:innen**
Ziel ist es, unsere entwickelte Server-Lösung in bestehende Qualitätssicherungsprozesse von Entwickler:innen großer Sprachmodelle (LLMs) zu integrieren. Wir werden das **KomMKonLLM**-

Framework weiter in einschlägigen Communities bewerben, insbesondere bei Fachveranstaltungen und Konferenzen wie z. B. Sec4Dev[8], um dessen Nutzen für die Qualitätssicherung beim Einsatz von LLMs aufzuzeigen.

2.) **LLM-Nutzer:innen**
Um LLM Nutzer:innen die Relevanz von Konsistenztests und die Fähigkeiten unseres **KomMKonLLM**-Frameworks zu demonstrieren planen wir die Teilnahme an Science-Communication Veranstaltungen, wie z.B. die *Lange Nacht der Forschung* (https://langenachtderforschung.at/).

3.) **Externe Forscher:innen**
Die generierten Tests sollen in Science Outreach-Aktivitäten demonstriert werden, um das Bewusstsein für potenzielle Risiken und Qualitätsaspekte von LLMs zu schärfen. Darunter fallen After-Work Events wir (Security-) MeetUps (https://www.sba-research.org/sba-meetup-groups/), aber auch wissenschaftliche Konferenzen, an denen wir im Zuge unserer Forschungsaktivitäten aktiv beitragen und teilnehmen.

4.) **Interne Forschungsaktivitäten**
Die Server-Lösung sowie die generierten Tests werden auch weiterhin in der Forschungsarbeit der MATRIS-Forschungsgruppe sowie in Kooperation mit internationalen akademischen Partnern eingesetzt.

Während der Arbeit an diesem Projekt haben wir einige Ideen für Weiterentwicklungen und Verbesserungen gesammelt, sei es durch eigene Beobachtungen oder durch Diskussionen im Rahmen unserer Outreach-Aktivitäten.
Viele User:innen von LLMs verwenden diese, um (zu versuchen) Probleme zu lösen, welche Fähigkeiten voraussetzen. Je nach LLM und Version kann das besser oder schlechter gelingen. Einem/Einer durchschnittlichen User:in ist oftmals nicht bewusst, welche logischen Fähigkeiten das verwendete LLM hat, bzw. welche integriert sind. Daher möchten wir Methoden zur Überprüfung logischer Schlussfähigkeit in LLMs entwickeln und das **KomMKonLLM**-Framework dadurch *erweitern (Erweiterung zum Testen der Logik-Fähigkeiten von LLMs*).

Außerdem haben wir einige Ansätze zur *Verbesserung* des **KomMKonLLM**-Frameworks gesammelt:

a. Nutzung der LLMs selbst zur Generierung semantischer Varianten (z. B. Synonyme), als Bestandteil einer erweiterten Testpipeline.

b. Multi-linguale Unterstützung, insbesondere die Erweiterung der Pipeline um deutschsprachige Module zur besseren Adaptierbarkeit auf nationale Bildungskontexte.

c. Verbesserte Interaktion mit der Testpipeline durch die Entwicklung einer intuitiven Benutzeroberfläche, die sowohl Eingabe als auch Analyse von Testsätzen ermöglicht.

Wir haben vor einige dieser Anknüpfungspunkte in einem Netidee-Folgeprojekt aufzugreifen.

[8] https://sec4dev.io/

10 Anregungen für Weiterentwicklungen durch Dritte

Über die oben erwähnten Punkte hinaus, eröffnen sich Weiterentwicklungsmöglichkeiten durch den allgemeinen Fortschritt von LLMs und deren engere Verflechtung in Nutzer:innen-Anwendungen.

Konkrete und detaillierte Anregungen für Entwickler:innen sind unter https://github.com/KomMKonLLM/KomMKonLLM/blob/main/HACKING.md zu finden.

Das **KomMKonLLM**-Framework ist so konzipiert, dass es sich flexibel erweitern lässt. Es ist also ein idealer Ausgangspunkt für Beiträge und Weiterentwicklungen durch Dritte. Potenzielle Anknüpfungspunkte liegen etwa in der Integration weiterer LLM-Backends, oder der Erweiterung des Testfall-Repositoriums um spezifische Anwendungsbereiche (z. B. juristische, medizinische oder mehrsprachige Testfälle). Auch die visuelle Aufbereitung der Ergebnisse lässt sich durch alternative Darstellungsformen oder interaktive Auswertungen weiter verbessern. Durch die offene Architektur und modulare Gestaltung können Entwickler:innen mit unterschiedlichsten Schwerpunkten das **KomMKonLLM**-Framework sinnvoll ergänzen und zur Weiterentwicklung robuster, transparenter LLM-Systeme beitragen.