



**Hochschule
Bonn-Rhein-Sieg**
University of Applied Sciences



**HOCHSCHULE
HAMM-LIPPSTADT**



Wissenschaftliche Gesamtauswertung der Chatprotokolle

Ein Kooperationsvorhaben empfohlen durch die:



INNOVATION DURCH KOOPERATION

Gefördert durch:

Ministerium für
Kultur und Wissenschaft
des Landes Nordrhein-Westfalen



Inhaltsverzeichnis

1. Einleitung	3
2. Methodik	3
3. Ergebnisse.....	4
3.1. Lösungsrichtigkeit	4
3.2. Fehlerart	4
3.3. Interpretation der Fehlerstruktur	5
3.4. Prompt-Typ	5
3.5. Umgang mit der MINT-Aufgabe.....	5
3.6. Aufgabenübernahme.....	5
3.7. Gesamt-Promptprofil	6
3.8. Nachfragen an den KI-Chatbot.....	6
3.9. Plausibilisierung/Reflexion	6
3.10. Nutzung von KI-Output	6
3.11. Metakognitive Aspekte.....	7
4. Diskussion	7
5. Fazit	8

Im ersten Auswertungsteil wurde ein Durchlauf an der Hochschule Bonn-Rhein-Sieg (HBRS) durchgeführt. Die Aufgabe 6 umfasste die Berechnung von Drehmomenten und Gleichstrommaschinen. Es wurden fünf Rückläufer berücksichtigt. Der zweite Auswertungsteil wurde an der Hochschule Hamm-Lippstadt (HSHL) durchgeführt. Die erste Aufgabe besteht in der Berechnung der Baseballbewegungen und Blocks auf dem Tisch. Die Anzahl der Rückläufer wird mit $n = 25$ angenommen. Daraus ergibt sich eine prozentuale Gewichtung von 16,66 Prozent für den Auswertungsteil 1 (HBRS) und 83,33 Prozent für den Auswertungsteil 2 (HSHL). Im ersten Auswertungsteil wurden 35 Datensätze analysiert, während im zweiten Auswertungsteil 125 Datensätze einer Auswertung unterzogen wurden.

1. Einleitung

Für die vorliegende Gesamtauswertung wurden die Datensätze der Hochschulen HBRS (Hochschule Bonn-Rhein-Sieg) und HSHL (Hochschule Hamm-Lippstadt) zusammengeführt. Die Stichprobe umfasst insgesamt 30 Rückläufer, in denen die Aufgabenteile 1 und 2 (HSHL) sowie Aufgabe 6 (HBRS) hinsichtlich verschiedener Merkmale analysiert wurden. Im Rahmen der Untersuchung wurden verschiedene Aspekte analysiert, darunter die Lösungsrichtigkeit, die Fehlerart, der Prompt-Typ, der Umgang mit der MINT-Aufgabe, die Aufgabenübernahme sowie das Gesamt-Promptprofil.

Zunächst wurde auf Grundlage der zusammengeführten Datensätze eine deskriptive Analyse der zentralen Merkmale durchgeführt. Der Untersuchungsschwerpunkt lag dabei insbesondere auf der Verteilung der Lösungsrichtigkeit, der auftretenden Fehlerarten sowie der Struktur der verwendeten Prompts.

Im Folgenden werden zunächst die Ergebnisse zur Lösungsrichtigkeit der bearbeiteten Aufgabenteile dargestellt. Im weiteren Verlauf werden die promptbezogenen Merkmale, darunter der Prompt-Typ, der Umgang mit der MINT-Aufgabe, die Aufgabenübernahme sowie das Gesamt-Promptprofil, einer näheren Betrachtung unterzogen, um mögliche Zusammenhänge zwischen Promptstruktur und Bearbeitungsergebnis zu identifizieren.

2. Methodik

Zunächst erfolgte die Kodierung der Chatprotokolle gemäß dem vorgegebenen Codierleitfaden. Im Anschluss daran wurde eine Überprüfung auf Lösungsrichtigkeit und Fehlerart durchgeführt. Die Daten wurden aus einer Pivot-Tabelle (Excel-Tabelle) extrahiert und als relative Häufigkeiten ausgewertet. Die vier Kompetenzstufen (L0–L3) repräsentieren unterschiedliche Grade der Lösungsrichtigkeit. Darüber hinaus wurden weitere Merkmale der Fehlerart und Promptstruktur analysiert, darunter der Prompt-Typ, der Umgang mit der MINT Aufgabe, die Aufgabenübernahme, Plausibilisierung/Reflexion sowie das Gesamt Promptprofil.

3. Ergebnisse

3.1. Lösungsrichtigkeit

Zur Bewertung der Lösungsrichtigkeit wurde ein Bewertungssystem von insgesamt vier Stufen entwickelt. Die Bewertung erfolgt nach folgendem Schema: L0 wird als "komplett falsch" klassifiziert, L1 als „teilweise richtig, mit wesentlichen Fehlern“, L2 als „überwiegend richtig, mit kleineren Fehlern“ und L3 als "vollständig korrekt".

Die Auswertung der Daten hinsichtlich der Bewertung der Lösungsrichtigkeit ergibt eine signifikante Konzentration bei der Lösungsrichtigkeit in beiden Auswertungsteilen bei L3 „vollständig korrekt“. Somit stellt dieser den größten Anteil dar. Eine Analyse der Daten ergibt, dass die höchste Kompetenzstufe L3 in beiden Datensätzen vorherrschend ist. Es konnte festgestellt werden, dass etwa zwei Drittel aller Lösungen das höchste Niveau erreichen.

Die Auswertung ergibt, dass der Anteil der Ergebnisse der Stufe L2 („überwiegend richtig, mit kleineren Fehlern“) im Auswertungsteil 2 (HSHL) mit 16,81 Prozent höher gewichtet ist als im Auswertungsteil 1 (HBRS), wo der Wert für die Stufe L2 bei 5,71 Prozent liegt. Die niedrigeren Kompetenzstufen im Teil 2 (HSHL) weisen mit einer Stichprobengröße von 25 eine deutlich geringere Inzidenz auf. Im Auswertungsteil 1 (HBRS) liegt der Anteil bei Stufe L1 „teilweise richtig, wesentliche Fehler“ bei 8,57 Prozent und im Auswertungsteil 2 (HSHL) bei 4,42 Prozent. Der Anteil von Level L0 ist im Auswertungsteil 2 (HSHL) mit 0,88 Prozent nahezu vernachlässigbar, wohingegen er im Auswertungsteil 1 (HBRS) 17,14 Prozent beträgt.

3.2. Fehlerart

Die vorliegende Fehlerart wurde anhand einer sechsstufigen Klassifikation codiert, wobei die Kategorien F1 „Rechenfehler“, F2 „Verständnisfehler“, F3 „Physikalisches Konzept falsch“, F4 „Halluzinationen/nicht zur Aufgabe passend“, F5 „Formelfehler/falscher Ansatz“ und F6 „Darstellungsfehler“ angewendet wurden.

Das Bewertungssystem, welches insgesamt sechs Stufen umfasst, wies bei der Auswertung des ersten Teils eine Verteilung bei der Fehlerart F1 „Rechenfehler“ von 41,67 Prozent auf. Die höchste Konzentration wurde hingegen im zweiten Teil (HSHL) mit 50 Prozent bei der Fehlerart F6 „Darstellungsfehler“ verzeichnet. Im Auswertungsteil 1 (HBRS) sind 33,33 Prozent der Fehler auf „Formelfehler/falscher Ansatz“ (F5) zurückzuführen, während bei den untersuchten Fällen an der HSHL Formfehler oder falsche Ansatzannahmen mit 17,85 Prozent identifiziert werden konnten. Im Auswertungsteil 2 der Kategorien F1 „Rechenfehler“ und F4 „Halluzinationen/nicht zur Aufgabe passend“ wurde jeweils eine gleiche Verteilung von 14,29 Prozent festgestellt. Die niedrigste Konzentration der Verteilung wurde dabei mit 3,57 Prozent bei F2 „Verständnisfehler“ (F2) verzeichnet. Die Auswertung ergab, dass die niedrigste Konzentration im ersten Auswertungsteil (HBRS) mit 25 Prozent bei „Physikalisches Konzept falsch“ lag. Die Fehlerarten F2 „Verständnisfehler“ und F4 „Halluzinationen/nicht zur Aufgabe passend“ traten nicht auf.

3.3. Interpretation der Fehlerstruktur

Die Ergebnisse zeigen, dass Fehler insgesamt relativ selten auftreten und sich vor allem auf die unteren Kompetenzstufen konzentrieren. Während L0 und L1 eher grundlegende Fehler oder unvollständige Lösungsansätze darstellen können, deutet die hohe Konzentration auf L3 darauf hin, dass die Mehrheit der Aufgaben vollständig oder nahezu vollständig korrekt gelöst wurde.

3.4. Prompt-Typ

Die Klassifizierung des Prompt-Typs erfolgt in die Kategorien P1a „Keine Interaktion sichtbar/nur eine Nachricht“, P1b „Nur Aufgabe reinkopiert, kein Prompting“, P2 „Minimal Prompting (z. B. 'Bitte Lösen')“, P3 „inhaltliches Prompting (z. B. Einschränkungen, Hinweise, Format)“ und P4 „Fortgeschrittenes Prompting (z. B. iterative Verbesserung, Nachfragen, Fehlerkorrekturen, CoT-Aufforderung)“.

Bei der Betrachtung des Prompt-Typs zeigt sich eine sehr hohe Homogenität. In beiden Datensätzen treten überwiegend identische Promptstrukturen auf. Diese starke Standardisierung der Prompts führt zu einer hohen Vergleichbarkeit der Ergebnisse, da Unterschiede in der Lösungsrichtigkeit weniger wahrscheinlich auf unterschiedliche Promptformulierungen zurückzuführen sind. Die Kategorie P1a tritt deutlich seltener auf und beschreibt Fälle, in denen keine Interaktion stattfindet. Die Analyse des untersuchten Datensatzes ergibt, dass überwiegend einfach strukturierte Promptformen Anwendung fanden, bei denen entweder lediglich die Aufgabenstellung übernommen wurde oder keine Interaktion stattfand. Die größte Verteilung lag mit 90,5 Prozent bei dem Prompt-Typ P1b „Nur Aufgabe reinkopiert, kein Prompting“ wobei der Prompt-Typ P1a bei 7,5 Prozent in der gesamten Betrachtung lag. Bei der Gesamtbetrachtung (n=30) lag die größte Verteilung mit 90,5 Prozent beim Prompt-Typ P1b „Nur Aufgabe reinkopiert, kein Prompting“. Der Prompt-Typ P1a lag bei 7,5 Prozent.

3.5. Umgang mit der MINT-Aufgabe

Auch beim Umgang mit der MINT-Aufgabe zeigt sich eine weitgehend einheitliche Struktur. Die Bearbeitung der Aufgaben erfolgt überwiegend nach einem vergleichbaren Muster. Dieses Ergebnis kann darauf hinweisen, dass die Aufgabenstellung klar formuliert war und von den Teilnehmenden in ähnlicher Weise interpretiert wurde.

3.6. Aufgabenübernahme

Die Kategorie der Aufgabenübernahme weist in beiden Auswertungsteilen auf eine nahezu vollständige Ausprägung hin. Die vorliegenden Beobachtungen legen nahe, dass die gestellte Aufgabe A0 „Aufgabe 1:1 reinkopiert per Screenshot/als Text/als Anhang (z. B. PDF)“ vollständig übernommen wurde. Im zweiten Auswertungsteil (HSHL) wurde eine weitere signifikante Ausprägung mit 11,2 Prozent bei A4 „Aufgabenteil vergessen reinkopieren“ festgestellt. Bei der Verteilung A1 „Aufgabe leicht bearbeitet/paraphrasiert“ liegt die Spannweite im Auswertungsteil 2 (HSHL) bei 0,8 Prozent. Die vorliegenden Beobachtungen legen nahe, dass die gestellte Aufgabe A0 „Aufgabe 1:1 reinkopiert per Screenshot/als Text/als Anhang (z. B. PDF)“ vollständig übernommen wurde oder auch teilweise, jedoch eher unerheblich vergessen wurde, reinkopiert zu werden. Darüber hinaus existieren weitere Kategorien, die wie folgt definiert werden können: Die Aufgabe wurde hinsichtlich A2 klar strukturiert, gekürzt und präzisiert und hinsichtlich A3

transformiert, d. h. in Einzelschritte zerlegt. Die Analyse ergibt, dass die Aufgabenübernahme in nahezu allen Fällen vollständig erfolgt. Die Teilnehmenden folgen somit in der Regel direkt der gegebenen Aufgabenstellung. Dies lässt den Schluss zu, dass die Impulsgeber prägnant formuliert waren und eine eindeutige Bearbeitungsrichtung vorgaben. In der Gesamtbetrachtung lag die Aufgabenübernahme bei mit einer signifikanten Mehrheit bei A0 „Aufgabe 1:1 reinkopiert per Screenshot/als Text/als Anhang (z. B. PDF)“ bei 90,63 Prozent. Eine niedrigere Konzentration konnte bei A4 „Aufgabenteil vergessen reinzukopieren“ mit 8,75 Prozent festgestellt werden. Eine nicht nennenswerte Konzentration von 0,63 Prozent war bei A1 „Aufgabe leicht bearbeitet/paraphrasiert“ erkennbar.

3.7. Gesamt-Promptprofil

Das Gesamt-Promptprofil zeigt ebenfalls eine vollständige Konzentration innerhalb der Kategorie G0 – Copy paste: „Aufgabe reinkopiert, keine Prüfung“. Ebenso ist das Gesamt Promptprofil durch eine ausgeprägte Homogenität gekennzeichnet. Obwohl dadurch lediglich eine begrenzte Variation zwischen verschiedenen Promptstrategien beobachtet werden kann, stellt diese Konsistenz dennoch ein auffälliges Ergebnis dar. Sie deutet darauf hin, dass sich ein relativ stabiles Interaktionsmuster zwischen Aufgabenstellung und Bearbeitung etabliert hat.

3.8. Nachfragen an den KI-Chatbot

Die Analyse der Interaktionen mit dem KI-Chatbot ergab, dass die Nachfragen an den KI Chatbot in beiden Auswertungsteilen (HSHL und HBRS) insgesamt mit dem Code R0 „Keine“ klassifiziert wurden. In diesem Kontext konnte eine Differenzierung zwischen den drei Typen von Nachfragen vorgenommen werden: R1 „Verständnisnachfragen“, R2 „Korrektur /Überprüfungsnachfragen“ und R3 „iterative Vertiefungsfragen“ differenziert werden.

3.9. Plausibilisierung/Reflexion

Im Hinblick auf die vollständige Konzentration wurde festgestellt, dass keine Plausibilisierung der Aufgaben erfolgte. Die Auswertung der Chatprotokolle (HBRS und HSHL) ergab für beide Ergebnisteile ein hundertprozentiges Ergebnis S0 („keine Reflexion, reine Übernahme der KI Antwort“).

3.10. Nutzung von KI-Output

Im Auswertungsteil 1 (HBRS) wies der Übernahmegrad insgesamt eine „vollständige Übernahme ohne Änderung“ (U0) auf. Es erfolgte weder eine leichte Modifikation (U1) noch eine starke Modifikation (U2). Auch wurde die KI nicht als Ideengeber genutzt welche als Codierung U3 „KI nur als Ideengeber genutzt, eigene Lösung erzeugt“ möglich gewesen wäre. Im Auswertungsteil 2 (HSHL) wurde mit 81,42 Prozent die Kategorie einer „vollständigen Übernahme ohne Änderung“ (U0) belegt. Darüber hinaus konnte eine leichte Modifikation (U1) mit einem Wert von 18,58 Prozent festgestellt werden. Es erfolgte keine signifikante Modifikation (U2) und die Nutzung der KI als Ideengeber war ebenfalls nicht realisiert (U3). Eine Codierung „KI nur als Ideengeber genutzt, eigene Lösung erzeugt“ (U3) war demnach nicht möglich.

In der Gesamtbetrachtung (n=30) war ein signifikantes Ergebnis bei U0 „vollständige Übernahme ohne Änderung“ mit 85,81 Prozent festzustellen. Eine niedrige Konzentration wurde bei U1 „leichte Modifikation“ mit 14,19 Prozent ermittelt.

3.11. Metakognitive Aspekte

In der Zusatzkategorie „Metakognitive Aspekte“ wurde das Bewusstsein der Studierenden evaluiert. Die Ergebnisse zeigen insgesamt ein einheitliches Bild: Es besteht ein fehlendes Bewusstsein (M1). Es erfolgt keine Überprüfung der Plausibilität und auch keine Prüfung der Lösungen, die nicht zuvor einer Prüfung unterzogen wurden. Des Weiteren wurden die Kategorien M0 – Bewusstsein: Plausibilisierung „Studierende fordern Überprüfung aktiv ein“ und M3 – Reflexion über Qualität der Chatbot-Antwort „Studierende kommentieren oder bewerten die Antwort“ berücksichtigt.

4. Diskussion

Die zusammengeführte Analyse zeigt eine insgesamt hohe Lösungsrichtigkeit innerhalb der gesamten Stichprobe. Die Mehrheit der Lösungen erreicht das höchste Kompetenzniveau (L3). Gleichzeitig zeigt sich eine starke Standardisierung in mehreren Promptmerkmalen, darunter der Prompt-Typ, der Umgang mit der Aufgabe und das Gesamt-Promptprofil. Die Ergebnisse der Gesamtauswertung zeigen insgesamt ein hohes Leistungsniveau der Teilnehmenden. In beiden Auswertungsteilen dominiert die höchste Kompetenzstufe L3 („vollständig korrekt“), was darauf hindeutet, dass die Mehrheit der Aufgaben mithilfe des KI-Chatbots erfolgreich bearbeitet werden konnte. Gleichzeitig treten niedrigere Kompetenzstufen (L0–L2) nur in geringem Umfang auf. Insbesondere im ersten Auswertungsteil wurde jedoch ein erhöhter Anteil an L0 beobachtet. Dies könnte auf Unterschiede im Kompetenzniveau, in den Lösungsstrategien oder in der Interpretation der Aufgabenstellung hindeuten.

Ein zentrales Ergebnis der Analyse ist die ausgeprägte Homogenität der Promptstrukturen. Der überwiegende Anteil der Teilnehmenden nutzte den Prompt-Typ P1b („nur Aufgabe reinkopiert, kein Prompting“). Komplexere Promptstrategien fanden hingegen kaum Anwendung. Diese Standardisierung setzt sich im Gesamt-Promptprofil (G0) sowie im Umgang mit der MINT-Aufgabe fort.

Darüber hinaus zeigt sich, dass die Aufgabenübernahme in den meisten Fällen vollständig und unverändert erfolgte (A0), mit nur geringfügigen Abweichungen durch unvollständige Übernahmen (A4) oder minimale Paraphrasierungen (A1). Die Analyse der Daten zeigt, dass der KI-Output überwiegend durch eine direkte Übernahme ohne Modifikation (U0) charakterisiert ist. Es finden sich nur wenige Anzeichen für eigenständige Anpassungen oder die Nutzung der KI als Ideengeber.

Auffällig ist zudem das Fehlen metakognitiver Aktivitäten: Es wurden weder Nachfragen an den Chatbot gestellt noch erfolgte eine Plausibilisierung oder Reflexion der Ergebnisse. Dieser Befund deutet darauf hin, dass die KI primär als unmittelbare Lösungsquelle und nicht als unterstützendes Werkzeug im Problemlöseprozess genutzt wurde.

Die Ergebnisse weisen darauf hin, dass die hohe Lösungsrichtigkeit weniger auf differenzierte Interaktionsstrategien, sondern vielmehr auf eine standardisierte Nutzung der KI in Kombination mit klar strukturierten Aufgabenstellungen zurückzuführen ist.

5. Fazit

Die Gesamtauswertung offenbart eine signifikant hohe Lösungsrichtigkeit bei gleichzeitig stark standardisierter Nutzung des KI-Chatbots. Die Mehrheit der Teilnehmenden erreichte das höchste Kompetenzniveau, wobei die Interaktion mit der KI überwiegend auf einfaches Kopieren der Aufgabenstellung und direkte Übernahme der Ergebnisse beschränkt blieb. Es konnte festgestellt werden, dass komplexere Promptstrategien sowie metakognitive Prozesse wie Reflexion oder Überprüfung nur in geringem Maße stattfanden.