„KI als neuer Mitarbeiter" – Wie Large Language Models Arbeitsprozesse transformieren

Christian Harms, merlin.zwo digital concept

Large Language Models (LLMs) wie GPT-4, LLaMA, Claude oder Mistral haben sich innerhalb kurzer Zeit zu leistungsfähigen Werkzeugen entwickelt, um textbasierte Prozesse im Unternehmenskontext neu zu denken. Ihr Mehrwert zeigt sich besonders dort, wo aus Freitext strukturierte Informationen generiert werden sollen – sei es in der Kundenkommunikation, der Verwaltung oder im medizinischen Bereich.

Was LLMs dabei von klassischen regelbasierten Ansätzen unterscheidet, ist nicht die Tatsache, dass sie Kategorien zuweisen oder Entscheidungen vorbereiten – sondern wie sie das tun: Sie arbeiten auf Basis eines tieferen Sprach- und Kontextverständnisses, nicht auf Grundlage starrer, manuell codierter Regeln. Die resultierende Klassifizierung wirkt dadurch nicht „intelligent" – sie ist es, weil sie auf sprachlich-logischen Zusammenhängen und semantischer Interpretation beruht.

In unseren Projekten ging es explizit darum, strukturierte Ausgaben zu erzeugen: E-Mails sollten nach Thema, Dringlichkeit und Stimmungslage eingeordnet werden. Medizinische Texte sollten konkrete Entitäten wie Diagnose, Maßnahme oder Fachbereich liefern. Das Ziel war also eine systematische, wiederholbare Klassifikation – jedoch auf einem Niveau, das klassische Automatisierung nicht leisten kann.

Die zentrale Steuerungskomponente dabei ist das *Prompt-Engineering.* Statt Logik in Code zu gießen, wird sie sprachlich modelliert. Anforderungen, Einschränkungen, Formate und Beispiele werden in einem Prompt formuliert, der das Modell präzise instruiert – und je nach Domäne iterativ geschärft werden muss.

Dieser Artikel stellt zwei praxiserprobte Szenarien vor:

Zum einen die automatisierte Vorverarbeitung von Support-Mails mit Sentiment-Analyse und zum anderen die Analyse medizinischer Akten mit einem lokal betriebenen LLM – inklusive datenschutzkonformer Architektur. Beide Anwendungsfälle zeigen, wie sich LLMs gezielt in produktive Workflows integrieren lassen, wenn man ihre Stärken richtig adressiert: Kontextverstehen, semantische Einordnung und flexible, aber kontrollierbare Ausgabeformate.

Praxisbeispiel 1: Kundenstimmung erkennen – Sentiment-Analyse im Supportpostfach

Ein eindrucksvolles Beispiel für den erfolgreichen Einsatz von KI in realen Geschäftsprozessen ist ein Projekt, das wir gemeinsam mit einem Kunden aus dem IT-Dienstleistungssektor als Proof-of-Concept (PoC) umgesetzt haben. Ziel war es, die Bearbeitung eingehender E-Mails im zentralen Supportpostfach durch den gezielten Einsatz von LLMs zu beschleunigen und qualitativ zu verbessern.

In klassischen Supportstrukturen ist das Postfach oft ein Nadelöhr: Jeden Tag treffen dort zahlreiche Anfragen ein, deren Relevanz und Dringlichkeit zunächst manuell eingeschätzt werden müssen. Die Folge: hohe Reaktionszeiten, unklare Prioritäten und mitunter übersehene kritische Eskalationen. Hier setzt unsere Lösung an – mit einer intelligenten Textanalyse, die automatisiert die inhaltlichen Eckpunkte einer E-Mail erkennt und die Stimmung des Absenders interpretiert.

Mithilfe eines sorgfältig entwickelten Prompt-Engineerings analysiert ein LLM jede eingehende Nachricht, beziehungsweise gegebenenfalls die gesamte vorhergegangene Korrespondenz und extrahiert dabei relevante Informationen wie:

- **Kategorie des Anliegens** (zum Beispiel technisches Problem, Anfrage, Beschwerde),
- **Priorität** (niedrig, mittel, hoch),

- **Status** (offen, gelöst, Rückfrage erforderlich),
- **Tonalität und Stimmung** (neutral, unzufrieden, verärgert).

Besonders spannend ist die automatische Einschätzung der Kundenzufriedenheit anhand sprachlicher Nuancen. So erkennt das Modell beispielsweise verschärfte Formulierungen, Wiederholungen von Beschwerden oder einen fordernden Ton – und kann darauf basierend Hinweise auf potenzielle Eskalationen geben. Diese Informationen werden in einem strukturierten Format an nachgelagerte Prozesse übergeben und helfen dem Support-Team dabei, gezielt und priorisiert zu reagieren.

Die Vorteile dieser Lösung sind bereits im PoC deutlich geworden:

- Reaktionszeiten konnten verkürzt werden, weil dringende Fälle automatisiert nach oben priorisiert wurden.
- Supportmitarbeiter erhielten eine fundierte Vorbewertung, ohne selbst lange Texte lesen zu müssen.
- Die Stimmungserkennung ermöglichte ein frühzeitiges Eingreifen bei drohender Unzufriedenheit.

Aktuell wird das System produktiv eingeführt und in bestehende Workflows integriert – mit dem Ziel, langfristig nicht nur die Effizienz zu steigern, sondern auch die Servicequalität messbar zu verbessern. Der Erfolg des Projekts zeigt: LLMs sind in der Lage, Sprache nicht nur zu verarbeiten, sondern wirklich zu verstehen – und das macht sie zu einem mächtigen Werkzeug im Kundenkontakt.

Praxisbeispiel 2: Strukturierte Informationen aus komplexen Dokumenten – KI in der medizinischen Aktenanalyse

Ein besonders spannender und anspruchsvoller Anwendungsfall für den Einsatz von Large Language Models ist die Analyse medizinischer Patientenakten. Gemeinsam mit einem Kunden aus dem Gesundheitswesen haben wir im Rahmen eines Proof-of-Concepts eine Lösung entwickelt, um medizinische Dokumente automatisiert zu analysieren und Informationen strukturiert zu extrahieren, sowie dem zuständigen Fachbereich zur Verfügung zu stellen.

Medizinische Texte wie OP-Berichte, Entlassungsschreiben oder klinische Verlaufsdokumentationen umfassen häufig Dutzende bis Hunderte von Seiten. Sie sind unstrukturiert, verwenden hochspezialisierte Fachsprache und enthalten kritische Informationen, die für nachgelagerte Prozesse – zum Beispiel Abrechnung, Qualitätssicherung oder medizinische Auswertung – zentral sind. Die manuelle Aufbereitung dieser Dokumente durch menschliche Mitarbeiter ist ressourcenintensiv und fehleranfällig.

Ziel unseres Ansatzes war es, mithilfe eines LLM aus solchen Dokumenten automatisch die wichtigsten Kopfdaten zu extrahieren – etwa die Fallnummer, Patientendaten, Diagnosen, durchgeführte Maßnahmen, Behandlungen und – insbesondere – daraus den zuständigen Fachbereich zu ermitteln. Die große Herausforderung bestand nicht nur im Erkennen dieser Inhalte, sondern in ihrer präzisen, strukturierten und validierbaren Aufbereitung, sodass sie anschließend in bestehende Prozesse und Systeme integriert werden können.

Datenschutz als zentrales Entscheidungskriterium

Von Beginn an war klar: In einem medizinischen Kontext steht der Schutz personenbezogener und besonders sensibler Gesundheitsdaten an oberster Stelle. Die Nutzung eines öffentlichen Online-Modells – etwa über eine Cloud-basierte API – war daher ausgeschlossen. Die Anforderung lautete: Die Daten dürfen das Unternehmen zu keinem Zeitpunkt verlassen.

Die Lösung: Der Einsatz eines lokal betriebenen LLMs, das innerhalb der eigenen IT-Infrastruktur betrieben wird. Dadurch bleiben alle Daten vollständig unter der Kontrolle des Unternehmens. Gleichzeitig ermöglicht dieses Setup maximale Flexibilität bei der Anpassung des Modells und bei der Integration in bestehende Sicherheitssysteme und Prozesse.

Natürlich bringt ein lokales Modell zusätzliche Herausforderungen mit sich – insbesondere hinsichtlich Ressourcenbedarf und technischer Komplexität. Doch der Datenschutz, sowie der Zugewinn an Datensouveränität und Compliance war für unseren Kunden ausschlaggebend. Zudem zeigt die Entwicklung im Bereich Open-Source-LLMs, dass leistungsfähige Modelle mittlerweile auch lokal einsetzbar sind – nahezu ohne Abstriche bei der Qualität.

Der Weg zum funktionierenden Prompt

Technisch gesehen waren Modelle wie LLaMA und Qwen out of-the-box in der Lage, medizinische Inhalte zu analysieren. Entgegen der Erwartung ein ressourcenintensives Finetuning durchführen zu müssen, konnten wir also direkt mit den Modellen arbeiten. Doch erst durch ein systematisch entwickeltes Prompt-Engineering wurden die Ergebnisse praxistauglich. Die Lernkurve dabei war steil. Es wurde schnell deutlich, dass ein erfolgreicher Einsatz nicht nur technisches Know-how, sondern auch fachliches Verständnis und viel Iteration erfordert.

Folglich flossen erhebliche Anstrengungen in die Entwicklung eines präzisen, robusten Prompts. Dabei waren fünf Faktoren entscheidend:

- natürlichsprachliche, klar formulierte Anforderungen,
- eine hohe Spezifität der Vorgaben,
- explizite Ausschlüsse möglicher Fehlerquellen,
- die Arbeit mit konkreten Beispieltexten und
- genaue Formatvorgaben für die Ausgabe.

Nach mehreren Optimierungszyklen konnte das Modell zuverlässig strukturierte Daten liefern – zum Beispiel zur Klassifizierung von Diagnosen, Zuordnung von Fachbereichen oder Erkennung von operativen Maßnahmen. Besonders hilfreich war die Möglichkeit, Negativbeispiele zu definieren und das Modell gezielt auf Ausnahmefälle vorzubereiten.

Die Sache mit der Kontextgröße

Die Verarbeitungskapazität von Large Language Models (LLMs) ist durch eine begrenzte Kontextlänge eingeschränkt. Es liegt auf der Hand, dass umfangreiche Dokumente, etwa hunderte Seiten lange PDFs, nicht vollständig und in einem Schritt an ein LLM übermittelt werden können. Konventionelle Ansätze wie Retrieval-Augmented-Generation (RAG),

die darauf abzielen, relevante Informationen aus großen Dokumenten mithilfe von LLMs zu extrahieren, stoßen hierbei schnell an ihre Grenzen. Dies zeigt sich insbesondere bei medizinischen Unterlagen, wie Patientenakten, in denen sich auf nahezu jeder Seite zahlreiche personenbezogene Angaben wie Namen und Adressen finden. Doch welche dieser Informationen gehören tatsächlich zum Patienten?

Zur Lösung dieses Problems wurde ein zweistufiges Vorgehen entwickelt: Zunächst wird das vollständige PDF seitenweise durch ein LLM analysiert, um eine strukturierte Übersicht der enthaltenen Informationen zu generieren. Auf Basis dieser Vorverarbeitung können anschließend gezielt die relevanten Bereiche identifiziert und dem LLM zur Extraktion spezifischer Inhalte – etwa des Patientennamens oder der Diagnosen – übergeben werden.

Das Modell der Wahl

Die Auswahl eines geeigneten Large Language Models geht weit über einen reinen Leistungsvergleich hinaus. Entscheidend ist ein fundiertes Verständnis der jeweiligen Stärken und Grenzen – etwa in Bezug auf Sprachverständnis, Genauigkeit und Kontextverarbeitung. Eine zentrale Rolle spielt zudem der Datenschutz: In sensiblen Anwendungsbereichen wie der Medizin oder dem Rechtswesen ist der Einsatz eines lokalen Modells oft unumgänglich, um regulatorische Anforderungen wie die DSGVO zuverlässig zu erfüllen.

Cloud-basierte LLMs – beispielsweise von OpenAI – bieten hohe Genauigkeit, kontinuierliche Weiterentwicklung und eine einfache Integration über standardisierte APIs. Gleichzeitig erfordern sie jedoch eine sorgfältige Prüfung hinsichtlich der Datensicherheit und -übermittlung, insbesondere bei personenbezogenen Informationen. Lokale Modelle wie LLaMA oder Qwen ermöglichen den Betrieb in vollständig kontrollierten Umgebungen und bieten maximale Datenhoheit. Sie setzen allerdings eine leistungsfähige Infrastruktur, technisches Know-how sowie einen erhöhten betrieblichen Aufwand voraus.

Grundsätzlich gilt: Je sensibler die Daten und je strenger die Compliance-Vorgaben, desto stärker spricht die Entscheidung für ein lokal betriebenes LLM. Bei allgemeinen Anwendungsfällen ohne besondere Datenschutzanforderungen überwiegen hingegen oft die Vorteile cloudbasierter Lösungen.

Der Schlüssel zur Qualität: Der richtige Prompt

Der Einsatz eines Large Language Models steht und fällt mit der Qualität der Anweisungen, die es erhält. Anders als bei klassischer Programmierung definiert man bei LLMs nicht einen festen Ablauf, sondern beschreibt möglichst präzise, was das Modell tun soll – und wie das Ergebnis aussehen soll. Dieses sogenannte *Prompt-Engineering* ist der zentrale Stellhebel für zuverlässige, konsistente und praxisnahe Resultate.

Im Rahmen unseres medizinischen Anwendungsfalls hat sich gezeigt, dass ein leistungsfähiger Prompt das Ergebnis systematischer Arbeit ist. Fünf Aspekte haben sich dabei als besonders erfolgskritisch erwiesen:

1. Natürlichsprachliche und strukturierte Anforderungen

Der Prompt muss in klarer, verständlicher Sprache formuliert sein – so, wie man auch einem Menschen eine Aufgabe erklären würde. Dabei sollte der Fokus auf Verständlichkeit und Eindeutigkeit liegen. Unklare Formulierungen oder mehrdeutige Begriffe führen schnell zu ungenauen Ergebnissen. Es hat sich als hilfreich erwiesen, die Anforderungen als Aufzählung zu formulieren. Am Beispiel der Patientenakte:

„Ermittle aus dem folgenden Text diese Informationen:

1. *„NAME_KLINIK“: Name der Klinik, in der der Patient behandelt worden ist.*
2. *„NAME_PATIENT“: Name des behandelten Patienten.*
3. *„ADRESSE_PATIENT“: Adresse des behandelten Patienten.*
4. *„GEBURTSDATUM_PATIENT“: Geburtsdatum des behandelten Patienten.*

2. Spezifität der Vorgaben

Je genauer der Prompt beschreibt, welche Informationen gesucht sind, desto besser sind die Resultate. Allgemeine Fragen wie „Was steht im Text?“ führen zu unbrauchbaren Ausgaben. Stattdessen sollte genau festgelegt werden, welche Inhalte extrahiert werden sollen – beispielsweise: *„Nenne die durchgeführte Maßnahme im medizinischen Eingriff und gib zusätzlich die zugehörige Diagnose an.“*

Wichtig sind zudem Vorgaben und Regeln, wo die Informationen zu finden sind. So befinden sich in einem Arztbrief Namen und Adressangaben im Briefkopf und in der Fußzeile, hierbei handelt es sich aber nicht um Informationen zum Patienten.

3. Explizites Ausschließen von Fehlern

Ein oft unterschätzter Faktor: Es reicht nicht, nur zu sagen, was das Modell tun soll, man muss auch klar benennen, was es nicht tun darf. Beispielsweise: „Gib keine Vermutungen ab“, „Vermeide Interpretationen bei unklaren Formulierungen“, oder „Wiederhole keine Inhalte, die bereits genannt wurden.“ Diese Negativbedingungen erhöhen die Konsistenz der Ausgaben erheblich.

4. Arbeiten mit Beispielen

LLMs arbeiten zwar auch allein mit expliziten Anweisungen sehr gut, jedoch profitiert die Qualität der Ergebnisse auch erheblich von Beispielen. Daher ist es hilfreich, dem Prompt konkrete Beispieltexte inklusive der gewünschten Ausgabeformate beizufügen. Auf diese Weise kann das Modell besser abstrahieren, was gemeint ist – und das Vorgehen auf andere Texte übertragen.

5. Formatvorgaben und Struktur

Ein weiterer kritischer Punkt ist das gewünschte Ausgabeformat. LLMs sind flexibel – manchmal zu flexibel. Wird nicht klar definiert, in welcher Struktur die Antwort zurückgegeben werden soll (zum Beispiel als JSON, Tabelle, Listenformat oder Klartext mit festen Überschriften), entstehen uneinheitliche Ausgaben, die nur schwer automatisiert weiterverarbeitet werden können. Beispiel:

„Erstelle ein JSON-Objekt mit folgenden Attributen: x, y, z, … Du darfst keine Anführungszeichen als Präfix oder Suffix verwenden und auch sonst keine einleitenden Worte in deiner Antwort ausgeben.“

Diese fünf Prinzipien waren das Ergebnis intensiver Arbeit mit dem Modell und

zahlreicher Testläufe. Mit dieser Vorgehensweise konnten wir verlässliche und reproduzierbare Ergebnisse erzielen – sowohl im medizinischen als auch im E-Mail-Szenario.

Fazit und Ausblick: KI als Mitgestalter der digitalen Arbeitswelt

Die beiden Praxisbeispiele zeigen eindrucksvoll, wie LLMs nicht nur theoretisch faszinieren, sondern ganz konkret Mehrwert im Arbeitsalltag schaffen. Ob im Supportbereich oder im Gesundheitswesen: Überall dort, wo Sprache verarbeitet, Informationen extrahiert und Entscheidungen vorbereitet werden müssen, kann KI zum produktiven Mitgestalter werden.

Dabei kommt es weniger auf die reine Rechenleistung an, sondern vielmehr auf die richtige Herangehensweise: Ein gut durchdachter Prompt ist oft wertvoller als ein noch leistungsfähigeres Modell. Die Erfahrung aus unseren Projekten zeigt, dass ein enger Austausch mit den Fachabteilungen, iterative Tests und das kontinuierliche Nachschärfen der Anforderungen essenziell sind, um von der beeindruckenden Leistungsfähigkeit der LLMs zu profitieren.

Der Blick in die Zukunft ist vielversprechend: Mit der wachsenden Verfügbarkeit lokaler LLM-Modelle, verbesserter Datenschutzmechanismen und leistungsstarker APIs rücken noch mehr Anwendungsfelder in greifbare Nähe. Gleichzeitig entstehen neue Herausforderungen – etwa bei der Validierung von Ergebnissen, der Integration in bestehende Systeme oder der Akzeptanz durch Anwender.

Doch eines ist bereits heute klar: KI ist kein Ersatz für menschliche Expertise – sie ist ein mächtiges Werkzeug, das Menschen entlastet, unterstützt und ihnen hilft, sich auf das Wesentliche zu konzentrieren. Und genau deshalb lohnt es sich, Künstliche Intelligenz als das zu begreifen, was sie zunehmend wird: ein neuer, wertvoller Mitarbeiter im digitalen Team.

Über den Autor

Christian Harms ist seit über 20 Jahren als Berater für Oracle-Datenbanken, Data-Warehouse-Systeme und Business Intelligence tätig. Sein aktueller Schwerpunkt liegt auf dem Einsatz Künstlicher Intelligenz und Large Language Models in Verbindung mit komplexen Datenbanksystemen, um bestehende Kundensysteme gezielt zu erweitern und manuelle Prozesse zu automatisieren.

Seine Erkenntnisse teilt er regelmäßig als Sprecher auf Fachkonferenzen und als Autor von Fachbeiträgen.

Christian Harms
christian.harms@merlin-zwo.de