Slide 1

Guten Tag, sehr geehrte/r Professor/in [Nachname Betreuer], sehr geehrte Anwesende. Mein Name ist [Dein Name] und ich freue mich, Ihnen heute die zentralen Erkenntnisse meiner Bachelorarbeit vorzustellen. In den kommenden 20 Minuten werden wir uns ansehen, wie die revolutionäre Transformer-Technologie die Fähigkeiten von Chatbots vorangetrieben hat und welche Herausforderungen sich daraus für die Praxis ergeben.

Slide 2

Hier sehen Sie den Fahrplan für die nächsten Minuten. Wir starten mit der Einleitung, um die Relevanz des Themas und das Forschungsproblem zu klären. Anschließend legen wir die theoretischen Grundlagen und vergleichen traditionelle NLP-Ansätze mit der Transformer-Architektur. Im Kern der Präsentation stelle ich Ihnen die drei wichtigsten Ergebnisse meiner Literaturanalyse vor. Diese werden wir danach gemeinsam einordnen und diskutieren, bevor ich die Arbeit mit einem Fazit und einem Ausblick auf zukünftige Forschung abschließe.

Slide 3

Wir alle kennen Chatbots. Aber in den letzten Jahren hat sich die Technologie dahinter fundamental geändert. Sogenannte Transformer-Modelle ermöglichen Konversationen, die sich fast menschlich anfühlen. Das ist die eine Seite der Medaille – das enorme Potenzial, das oft gefeiert wird. Auf der anderen Seite stehen jedoch die realen Hürden: immense Kosten, eine extreme Abhängigkeit von Daten und ethische Risiken. Genau in diesem Spannungsfeld – der Lücke zwischen Hype und Realität – setzt meine Arbeit an. Denn es fehlt bisher eine strukturierte Analyse, die nicht nur den Leistungssprung erklärt, sondern auch die damit verbundenen Herausforderungen ganzheitlich beleuchtet. Daraus ergibt sich die zentrale Frage meiner Arbeit...

Slide 4

Aus der Forschungslücke leitet sich direkt unsere zentrale Forschungsfrage ab: In welchem Maße prägen Transformer-Modelle eigentlich die Entwicklung und die Leistung von heutigen Chatbots? Um diese umfassend zu beantworten, habe ich mir drei konkrete Ziele gesetzt. Erstens: eine genaue Analyse der Technik. Was macht Transformer so leistungsfähig? Zweitens: der kritische Vergleich. Wo genau liegen die Vorteile, aber auch die Nachteile gegenüber den älteren Modellen? Und drittens: die Bewertung der Konsequenzen. Denn hohe Leistung hat ihren Preis – sei es bei den Kosten, der Datenqualität oder ethischen Fragen. Diese drei Ziele bilden das Gerüst meiner Analyse. Um sie zu erreichen, müssen wir zunächst die technologischen Grundlagen verstehen.

Slide 5

Um zu verstehen, warum Transformer so ein Durchbruch sind, müssen wir kurz auf ihre Vorgänger schauen: die Rekurrenten Neuronalen Netze oder kurz RNNs. Stellen Sie sich vor, Sie lesen einen Text, aber immer nur ein Wort auf einmal, von Anfang bis Ende. Genau das macht ein RNN. Wie dieses Diagramm zeigt, wird alles sequenziell, also Schritt für Schritt, verarbeitet. Das führt zu drei zentralen Problemen: Erstens: Es ist langsam. Man kann die Verarbeitung nicht parallelisieren, also beschleunigen. Zweitens, und das ist das größte Problem: RNNs haben ein 'kurzes Gedächtnis'. Bei langen Sätzen oder Dialogen 'vergessen' sie, was am Anfang gesagt wurde. Für einen Chatbot, der einen Gesprächsfaden halten muss, ist das fatal. Drittens sind sie ineffizient. Das Training dauert lange, was sie für die riesigen Datenmengen, die heute üblich sind, unpraktikabel macht. Diese fundamentalen Schwächen waren der Grund, warum die Forschung nach einer besseren Lösung suchte. Und diese Lösung...

Slide 6

...ist die Transformer-Architektur. Sie wurde speziell entwickelt, um die 'Vergesslichkeit' der RNNs zu überwinden. Der Schlüssel dazu ist ein völlig anderes Verarbeitungsprinzip, das Sie hier schematisch sehen. Statt Wort für Wort durch einen Satz zu gehen, schaut der Transformer sozusagen auf den gesamten Satz auf einmal. Jedes Wort wird in Beziehung zu jedem anderen Wort gesetzt. Das wird durch zwei entscheidende Innovationen ermöglicht: Erstens, der Self-Attention-Mechanismus. Das Modell 'lernt', welche Wörter in einem Satz für das Verständnis eines bestimmten Wortes am wichtigsten sind. Dadurch verliert es den Kontext auch bei langen Sätzen nicht aus den Augen. Zweitens, die parallele Verarbeitung. Da nicht mehr Schritt für Schritt gerechnet werden muss, kann die gesamte Aufgabe auf leistungsstarker Hardware parallelisiert werden. Das macht die Modelle extrem schnell und skalierbar. Und drittens führt das zu einem überlegenen Kontextverständnis. Durch einen Trick namens 'Multi-Head Attention' kann das Modell Mehrdeutigkeiten, wie zum Beispiel das Wort 'Bank', je nach Kontext korrekt interpretieren. Zusammenfassend lässt sich sagen: Diese neue Architektur ist der technische Grund für den Quantensprung in der Leistungsfähigkeit. Sehen wir uns nun an, was das für konkrete Ergebnisse in der Literatur bedeutet.

Slide 7

Kommen wir nun zum ersten von drei Schlüsselergebnissen meiner Analyse. Die Literatur zeigt ein sehr klares Bild: Transformer sind technologisch überlegen. Wie Sie hier an den konkreten Zahlen aus zitierten Studien sehen, erreichen sie in verschiedenen Aufgaben eine signifikant höhere Leistung als ihre Vorgänger. Aber es geht nicht nur um technische Benchmarks. Wichtiger für die Praxis ist: Dieses bessere Kontextverständnis führt zu natürlicheren Dialogen. Eine Studie konnte sogar eine um über 6 % höhere Nutzerzufriedenheit nachweisen. Der Grund dafür ist, wie wir eben gehört haben, der Self-Attention-Mechanismus, der die fundamentalen Schwächen der alten Modelle überwindet. Es handelt sich also nicht um eine kleine Verbesserung, sondern um einen echten Paradigmenwechsel. Doch diese Leistung hat ihren Preis, wie wir gleich sehen werden.

Slide 8

Überleitung von der letzten Folie: 'Aber diese beeindruckende Leistung hat, wie gesagt, ihren Preis...'. Erläutern Sie die drei Säulen des Ressourcenbedarfs anhand der Grafik: Hardware, Energie und Daten. Konkretisieren Sie die Zahlen: '600 bis 2.000 Euro im Monat sind für ein kleines Unternehmen eine massive Hürde'. Betonen Sie die ökologische Dimension – das ist nicht nur ein Kostenfaktor, sondern auch eine gesellschaftliche Verantwortung. Die Kernaussage am Ende scharf formulieren: Diese Hürden führen zu einer Machtkonzentration bei wenigen großen Playern und schränken den Zugang zur Technologie ein. Nächste Folie: 'Doch selbst wenn man die Ressourcen hat, gibt es eine weitere, noch kritischere Abhängigkeit: die der Datenqualität...'

Slide 9

Nachdem wir über Leistung und Kosten gesprochen haben, kommen wir zum dritten, vielleicht kritischsten Ergebnis: der Abhängigkeit von Daten. Das Prinzip ist einfach und als 'Garbage in, Garbage out' bekannt: Die Qualität der Ausgabe hängt direkt von der Qualität der Eingabe ab. Konkret bedeutet das zwei Hauptprobleme: Erstens: 'Halluzinationen'. Das sind keine Visionen, sondern faktisch falsche, aber überzeugend formulierte Aussagen. Ein Chatbot könnte zum Beispiel eine nicht existierende Studie zitieren. Das ist ein enormes Risiko, gerade in der Medizin oder Rechtsberatung. Zweitens: Bias. Wenn ein Modell hauptsächlich mit Texten aus einer bestimmten Kultur oder mit historischen Daten trainiert wird, die Vorurteile enthalten, lernt und reproduziert es diese. Das kann zu diskriminierenden Ergebnissen führen. Das Problem ist riesig, denn die Datenmengen sind gewaltig. Die Arbeit nennt das Beispiel OpenGPT-X: Von 2,5 Billionen Tokens wurden 40% allein durch Filtern entfernt. Die Sicherstellung der Datenqualität ist also eine immense Herausforderung. Diese Abhängigkeit ist nicht nur ein technisches, sondern ein tiefgreifendes ethisches Problem und die größte Hürde für den verlässlichen Einsatz. Nun, da wir diese drei zentralen Ergebnisse haben – die technologische Überlegenheit, die hohen Kosten und die Datenabhängigkeit –, wollen wir sie im nächsten Schritt einordnen und diskutieren.

Slide 10

Nachdem wir die drei Schlüsselergebnisse gesehen haben, ordnen wir diese nun ein. Es ist entscheidend zu verstehen, warum diese Effekte auftreten. Erstens: Die überragende Leistung ist kein Zufallsprodukt, sondern eine direkte Folge des Designs. Die Transformer-Architektur wurde exakt dafür gebaut, die 'Vergesslichkeit' der alten Modelle zu lösen. Das stützt die Theorie, dass die Architektur fundamental überlegen ist. Zweitens: Die enormen Kosten sind ebenfalls kein Zufall. Sie sind die direkte, theoretisch begründete Konsequenz aus der Komplexität dieser Modelle. Wenn ein Modell wie GPT-3 hunderte Milliarden Parameter hat, braucht es zwangsläufig immense Ressourcen. Es ist eine inhärente Eigenschaft des Modells. Dritter und wichtigster Punkt: Die Probleme mit Halluzinationen und Bias sind keine Schwäche der Transformer-Architektur an sich. Es ist eine Schwäche des gesamten datengetriebenen Ansatzes. Jedes Modell, das nur aus Daten lernt, wird durch die Qualität dieser Daten begrenzt. Dieser Befund ordnet das Problem also eine Ebene höher ein – es ist eine fundamentale Herausforderung des maschinellen Lernens. Diese Einordnung hilft uns, die Limitationen, die wir uns als Nächstes ansehen, besser zu verstehen.

Slide 11

Nachdem wir die Ergebnisse eingeordnet haben, fassen wir nun die zentralen Limitationen zusammen, die sich daraus ergeben. Erstens, die technische Grenze: Stellen Sie sich ein Gespräch vor, das nach 20 Seiten plötzlich den Anfang vergisst. Das ist die Realität der fixen Kontextfenster. Sie begrenzen, wie viel Information ein Modell gleichzeitig verarbeiten kann. Zweitens, die praktische Hürde: Wie wir schon bei den Ergebnissen gesehen haben, erfordert hohe Leistung massive Investitionen in Hardware und Energie. Das schafft eine Kostenbarriere, die den Zugang zur Technologie einschränkt. Drittens, und das ist die vielleicht gefährlichste Hürde: die qualitative Grenze. Die Modelle können faktisch falsche, aber plausibel klingende Antworten generieren. Das Vertrauen in die Ergebnisse ist also immer mit Vorsicht zu genießen. Diese drei zentralen Herausforderungen führen uns direkt zur Schlussfolgerung meiner Arbeit...

Slide 12

Wir sind am Ende unserer Analyse angelangt. Fassen wir die zentralen Erkenntnisse zusammen. Meine Arbeit kommt zu dem Schluss, dass Transformer-Modelle einen echten Paradigmenwechsel darstellen. Sie haben die Qualität und die Fähigkeiten von Chatbots revolutioniert, wie die linke Seite der Waage zeigt. Gleichzeitig – und das ist die entscheidende Erkenntnis – hat dieser Fortschritt seinen Preis. Auf der rechten Seite sehen wir die systemischen Herausforderungen: immense Kosten, eine grundsätzliche Unzuverlässigkeit durch Halluzinationen und tiefgreifende ethische Fragen. Wichtig ist hierbei die Erkenntnis: Diese Probleme sind keine Kinderkrankheiten, die einfach verschwinden, sondern sie sind inhärenter Teil des aktuellen technologischen Ansatzes. Der wissenschaftliche Beitrag meiner Arbeit ist es, genau dieses komplexe Spannungsfeld klar und strukturiert aufzuzeigen. Sie liefert damit eine Art Landkarte, die hilft, die oft überhitzte Debatte zu versachlichen und eine Grundlage für zukünftige Entwicklungen zu schaffen. Doch bei dieser Feststellung wollen wir nicht stehen bleiben. Aus dieser Analyse ergeben sich klare Empfehlungen für die Zukunft...

Slide 13

Wie auf der letzten Folie angekündigt, wollen wir nicht bei den Problemen stehen bleiben, sondern nach vorne schauen. Aus meiner Analyse ergeben sich drei klare Forschungsempfehlungen. Erstens, Effizienz und Nachhaltigkeit: Wir brauchen cleverere, hybride Modelle, um die enormen Kosten und den Energieverbrauch zu senken. Das ist die direkte Antwort auf die Ressourcen-Hürde. Zweitens, Ethik und Akzeptanz: Technik allein reicht nicht. Wir brauchen interdisziplinäre Studien, die den Menschen in den Mittelpunkt stellen, um Vertrauen aufzubauen und Bias abzubauen. Und drittens, Robustheit: Wir müssen die Zuverlässigkeit der Modelle verbessern. Dafür benötigen wir bessere Messmethoden, die uns helfen, das Problem der Halluzinationen systematisch anzugehen. Zusammengenommen bedeutet das: Der Weg in die Zukunft ist kein rein technischer, sondern ein ganzheitlicher, der technische, soziale und ethische Aspekte verbinden muss. Damit bin ich am Ende meiner Präsentation angelangt.

Slide 14

Damit bin ich am Ende meiner Präsentation angelangt. Ich bedanke mich herzlich für Ihre Aufmerksamkeit und insbesondere bei [Name des Betreuers] für die hervorragende Betreuung dieser Arbeit. Nun freue ich mich sehr auf Ihre Fragen und Anmerkungen.

Slide 15

Diese Folie nur bei direkter Nachfrage zu den konkreten Leistungsdaten zeigen, z.B. wenn jemand fragt: 'Können Sie die Überlegenheit quantifizieren?'. Erklären Sie kurz die drei unterschiedlichen Aufgaben, um die Vielseitigkeit zu betonen. Heben Sie hervor, dass die Verbesserung konsistent über verschiedene Aufgaben und Metriken hinweg auftritt. Das untermauert das Argument des Paradigmenwechsels.