Bullshitting LLMs and harmful intent: How to stay in control?

30. September 2023

Friedrich Answin Daniel Motz
daniel.motz@uni-jena.de
Friedrich-Schiller-Universität Jena

Einführung

Large Language Models (LLM) werden bereits heute von 100 Millionen Menschen täglich genutzt (Nutzer von ChatGPT lt. OpenAI). Es stellt sich nun die Fragen (ohne Anspruch auf Vollständigkeit):

- Welche Gefahren gehen von LLMs aus?
- Welche Gesetze sind notwendig um Gefahren abzuwenden?
- Wie stellt man die Einhaltung der Regelungen bei großen Herstellern, geschäftlichen und privaten Anwendern und Hobby-LLM-Trainern sicher?
- Können Moralvorstellungen einem KI-System nähergebracht werden?

Um die Realisierbarkeit einzuschätzen, ist es daher unabdinglich sich mit den Fähigkeiten und Limitationen von LLMs auseinanderzusetzen.

Terminologie

In dieser Arbeit werden antropomorphe Begriffe wie „lernen“, „verstehen“ und „entscheiden“ verwendet. Diese werden aufgrund der gängigen Terminologie im Englischen ins Deutsche übernommen. Sie sollen keinesfalls den Eindruck eines im menschlichen Sinne intelligenten Systems erwecken.

Fähigkeiten von LLMs

LLMs können an diverse symbolische Systeme angebunden und so eine Steuerzentrale für verknüpfte Anwendungen (Task Pipelines) sein. ChatGPT kann etwa mit einem Interpreter oder Compiler verknüpft, die Korrektheit eines Programmes überprüfen.

Weiterhin soll ChatGPT (genauer „GPT-4 early“) mutmaßlich einige Rezepte für gefährliche Chemikalien und Angriffsmöglichkeiten für terroristische Anschläge wiedergeben können. Es informiert über illegale und legale Kaufmöglichkeiten von Schusswaffen, gibt Rat für einen Mordanschlag, schreibt einen Massenvergewaltigungsdrohbrief und wägt Selbstverletzungsmethoden für den Nutzer ab. OpenAI zeigt jedoch, dass GPT-4 mit Fine-Tuning (genannt „launch“) eine Antwort auf derartige Prompts verweigert. [1]

Die potentiellen Möglichkeiten sind noch nicht absehbar. Ermöglicht es vorher mittellosen Akteuren den Einstieg in die Kriminalität? Kann man den öffentlichen Diskurs in sozialen Medien damit (besser) steuern?

AI Safety und Alignment

Eine Teildisziplin der KI-Forschung ist, Mechanismen zum Kontrollieren der Einhaltung der Ziele des Menschen im Verhalten von KI-Systemen zu überprüfen. Dies beinhaltet etwa Moral, Wertschätzung des Lebens und die zu lösende Aufgabe. Insbesondere eine Artificial General Intelligence (AGI) fordert Überprüfungsmechanismen heraus: sofern sie dem Anspruch genügt sollte die AGI imstande sein Mechanismen umgehen zu können.

Das Alignment-Problem ist mehrschichtig – die Regeln müssen exakt formuliert und mit einer Belohnung im Optimierungsverfahren (OV) assoziiert sein. Zweitens muss das OV trotz Regeln weiterhin die Ziele voranbringen – man spricht im negativen Falle von Reward Hacking. Und sollte es sich um eine AGI handeln – verstellt sie sich um den Entwickler in Sicherheit zu wägen?

Bias

Maschinelle Lernverfahren stellen aus Informationen mit einem Bias unerwünschte Assoziationen her – sind in den Trainingsdaten etwa überwiegend Hunde Chihuahuas, so wird aufgrund dieses Zusammenhangs auf die Frage

nach Hunden die Idee eines Chihuahuas wiedergegeben.

Outer Alignment (OA)

Um bestimmte Aufgaben lösen zu können, ist eine korrekte, vollständige Beschreibung notwendig. Hier ergibt sich das erste Problem: in wessen Interesse arbeitet ein KI-System? Die möglichen Parteien (die nicht mal untereinander einer Auffassung sein müssen) seien hier (erneut ohne Anspruch auf Vollständigkeit): die Entwickler, die Benutzer, der Gesetzgeber, die Wertegemeinschaft und die Moralisten. Selbst unter der Annahme, es seien die gewünschten Ziele eindeutig niedergeschrieben, müssen diese nicht die erwünschten Auswirkungen besitzen. Jede Maßnahme könnte gar unvorhersehbare Folgen (vgl. Goodharts Gesetz) besitzen.

Microsoft Chatbot „Tay" wurde im Jahr 2016 auf Twitter mit dem Ziel freigesetzt, vom Verhalten der Nutzer zu lernen. Es wurde, um die Erwartungen niedrig zu halten, mit der Persönlichkeit eines Teenagers ausgestattet. Die Intention der Designer eine im Alltag selbstlernende KI-Lösung zu bauen endete in mitunter antisemitischen und rassistischen Äußerungen des Chatbots. Eine wichtige Lektion für AI-Safety: von seinem Umfeld zu lernen (man könnte auch manipulierbar zu sein sagen) ist kein optimaler Ansatz. Den Designern war insbesondere nicht bewusst: ein KI-System „denkt" nicht – vor reflektiert es nicht, ob ein Verhalten wünschenswert ist oder moralischen Grundsätzen entspricht. Es folgt dem, was ein Algorithmus vorschreibt.

Inner Alignment (IA)

Hier gilt es zu prüfen, ob das formulierte Ziel tatsächlich gelernt wurde oder ob es nur den Anschein erweckt:

Train-Test-Splitting teilt den ursprünglichen Datensatz (mit bekannten Features und Labels) in bspw. 80% Trainings-Daten und 20% Verifikationsdaten. Nach dem Training können die Testfeatures zur Inferenz eingegeben und die vorhergesagten Labels mit den Testlabels abgeglichen werden. Analog kann dies mit den Trainingsdaten geschehen. Die resultierenden Train- und Test-Genauigkeit gibt Aufschluss darüber, ob die Trainingsdaten von einem zu komplexen Modell „auswendig gelernt" wurden (Overfitting, $\mathrm{acc}_{\mathrm{train}} > \mathrm{acc}_{\mathrm{test}}$) oder die Komplexität des Modells nicht ausreicht (Underfitting, beide Genauigkeiten niedrig).

Diversified Testing (kein Fachbegriff) meint, dass zur Identifikation von Misalignment möglichst divers getestet werden muss. Trainingsdaten können Biases (oder bestimmte Konstellationen gar nicht) enthalten. Sollte etwa ein KI-System trainiert werden ein Labyrinth lösen und wären diese Ziele in den Trainingsdaten immer grün, so besteht die Möglichkeit fälschlicherweise nach grünen Objekten statt etwa einer bestimmten Form, die die Ziele ebenfalls besitzen, zu suchen. Dies kann nur durch besseres Testen entdeckt und nur durch mehr Trainingsdaten umgangen werden. Grundsätzlich verhindern kann man dies im maschinellen Lernen allerdings nicht.

Ungelöst sind bisher Probleme wie Halluzination in LLMs. Es ist gleichzeitig in gewissem Maße notwendig zur Generation von interessanten Texten. Gleichzeitig geben LLMs mit bestechender Souveranität konstruierte „Fakten" wieder. In fiktionalen Texten womöglich erwünscht, schadet es in sachlichen Ausarbeitungen der Gesamtglaubwürdigkeit und bedingt eine menschliche Prüfung.

Misalignment

Large Language Models wie GPT-4 lehnen erst nach entsprechendem Fine-Tuning mit expliziten Beispielen (die die OpenAI Content Policy [2] abdecken sollen) schädliche Prompts ab, trotz der Vorfilterung des Trainingskorpus. Der Technical Report zeigt jedoch, dass auch GPT-4 Launch durch das Fine-Tuning nicht in allen Fällen die Antwort ablehnt [1]. Jailbreaks waren in GPT-3 und -4 lange nach Launch noch möglich (und neue Umgehungen werden weiterhin bekannt). Anhand der gleichartigen Antworten auf gegen die Content Policy verstoßende Prompts ist der Schluss naheliegend, dass ein weiterer Klassifzierer die generierten Antworten und Prompts auf Konformität mit OpenAIs Content Policy prüft.

Selbst mit restriktivem Vorgehen sind unerwünschte Äußerungen nicht abstellbar. Der

Einsatz von LLMs im Umgang mit Minderjährigen in der Bildung ist daher völlig undenkbar. Werden die Antworten weiters nicht auf inhaltliche Korrektheit überprüft bzw. Halluzinationen abgestellt, ist es zur Bildung für Kinder prinzipiell ungeeignet.

Eher vorstellbar ist es als natürlichsprachliches Interface für reguläre Anwendungen (Chat-Bot eines Consumer-Dienstleisters). Man lässt die Nutzerantwort durch das LLM „interpretieren", welches dann eine Aufforderung an bestehende Dialogprogramme sendet. So bestünde keine Gefahr eine obszöne Antwort nach außen geraten zu lassen. Dieses System sollte jedoch so gestaltet sein, dass keine Manipulationsmöglichkeiten bestehen. Jailbreaks und unvorhersehbare Antworten sind möglich und veranlassen die Einrichtung von menschlichen Konstrollinstanzen.

Bullshitting

… ist ein von Harry G. Frankfurt geprägter Begriff, der so viel bedeutet wie „die Absicht ohne Beachtung der Wahrheit zu überzeugen" (übersetzt auf dem Englischen).

LLMs können aufgrund ihrer Architektur keinen Quellennachweis liefern. Schlimmer noch: eine Frage, die sich der „Kenntnis" eines LLMs entzieht, wird durch schlichtes Generieren von Wörtern basierend auf Assoziationen beantwortet – das ist alles was die Architektur Transformer bieten kann. Spätestens nach Abschluss des Trainings, sind alle Bezüge zum Korpus nicht wiederherstellbar. [3]

Hier scheint eine erklärbare, symbolische Architektur notwendig. Ansätze in Explainable AI sind jedoch bisher nur für Bildklassifikation entwickelt und diese nur als post-hoc Erklärungen; für Verfahren die an sich erklärbar sind fehlt die entsprechende Flexibilität um sie auf LLMs anzupassen [4].

Was ist das Ziel von ChatGPT?

Es soll möglichst überzeugende, dialogähnliche Konversationen führen. Der Anspruch ist, dass auf jegliche Fragen mit einer sinnvollen Antwort bedient werden. Dabei sollen die ethischen Regeln von OpenAI (content policy) eingehalten werden: unter anderem verbietet sie Gewaltverherrlichung, Hassrede und Anregung zur Gewalt oder Straftaten.

Das Training und Fine-Tuning auf zufriedenstellende Antworten geschah mithilfe von Reinforcement Learning from Human Feedback (RLHF), realisiert durch ein Policy Model, das auf Rankings von Menschen basiert. Diese Rankings wurden allerdings nach dem Prinzip „preferred by human labelers" [1] – dies impliziert keinen Anspruch auf inhaltliche Korrektheit.

Malicious Intent

Die Ausgaben eines LLMs sind durch

- den Textkorpus des Grundmodells,
- Fine-Tuning (bspw. mithilfe von RLHF oder Prompt-Antwort-Beispielen von anderen LLMs),
- verfügbaren Rechner (der die Parametergröße des LLMs, und damit seine Qualität, limitiert) und
- Zugriff des Anwenders (sind weitere Filter für unerwünschte Prompts und Antworten eingebaut, vgl. ChatGPT)

limitiert und bedingt.

Jeder am Erstellen und Verwenden eine LLMs Beteiligte kann böse Absichten besitzen. Im Kontext von LLMs gibt es drei Parteien.

Sprachmodell

Das LLM an sich besitzt keine Intention. Es generiert lediglich Texte, die im Trainingskorpus assoziiert waren. Dennoch können von seinem Designer unerwünschte Antworten vorkommen; trotz Fine-Tuning. Es besteht daher keine Möglichkeit einem LLM ohne äußere Filter den Verstoß gegen die Regeln, einen „eigenen Willen", auszutreiben.

Nutzer / Privatpersonen

Sie sind durch die Modelle und ihre technische Versiertheit, die ihnen zur Verfügung stehen limitiert. Hinsichtlich LLMs, die 7, 40 oder 100 Milliarden Parameter haben können, ist bereits die Inferenz schwierig – aber ohnehin ist das Training nur für entsprechend finanziell ausgestattete und fachlich gebildete Personen möglich.

Nach aktuellem Stand ist das Training nicht ohne Einsatz eigener Ressourcen möglich (bspw. durch Probeversionen von Rechenzentren).

Designer / Entwickler

Zum Training eines LLMs sind beachtliche Ressourcen (bspw. 384 Nvidia A100 im Falle von Falcon-40B [5]) notwendig. Die Zahl an Organisationen mit dem Interesse diese Mittel aufzuwenden hält sich in Grenzen; damit auch der Aufwand diese zu regluieren. Open-Source LLMs wie Falcon können jedoch mithilfe von Fine-Tuning zu einem GPT-4-early-ähnlichen Zustand gebracht werden: hier könnten wiederum einige wenige Privatpersonen die Modelle zu ihren Zwecken abrichten. Sollten also die Quellen von LLMs nicht publiziert werden dürfen? Wie bereits festgestellt ist es nicht möglich einen hinreichenden Filter nur in den Gewichten des neuronalen Netzes einzubauen. Sollten also nur Berechtigte, etwa der Staat, Zugriff auf diese Technologie haben?

Reale Einsatzmöglichkeiten

… sind:

- Generation von Phishing-Nachrichten
- Shitstorms und Mobbing
- Verbreitung / Inflation von Agenden / Meinungen in sozialen Medien
- Generation von Nachrichten / Blog-Einträgen
- Automatisierte Video Pipeline

Selbstexperiment 1: Die Tauglichkeit von frei verfügbaren LLMs wurde beispielhaft am bereits erwähnten Falcon in seiner Instruct-7B-Parameter-Ausführung evaluiert.

Die Installation und Inbetriebnahme von Falcon gestaltete sich dank der verständlichen Anleitung auf *huggingface.co* leicht. Zur Inferenz war jedoch eine zur Nvidia RTX 2080 vergleichbare Grafikkarte notwendig.

Die Generation schädlicher Inhalte gestaltete sich schwierig. Beim Fine-Tuning von Falcon-Instruct verwendete man unter anderem OpenAI-GPT generierte Prompt-Antwort-Paare. Zum Prompting nutzte ich Beispiele aus der GPT-4 System Card. Etwa 10% meiner Prompts lieferten schädliche Ergebnisse.

Selbstexperiment 2: Mit einer „Video-Pipeline“ ist es möglich Video-Ideen, Skripte, Voice-Over und Video-Schnitt voll zu automatisieren. Hierzu bedarf es nur eines Zugangs zur OpenAI API, Google API, eines von ChatGPT generierten Skripts zum hochladen der Videos und ein wenig Zeit zum integrieren von *„Rytr.me“*, *„elevenlabs.io“* und *„pictory.ai“*.

Beispielhaft ließ ich nur ein Video generieren, was mich manuell ungefähr 25 Minuten inklusive Recherche kostete. Das größte Problem war das korrekte Formulieren eines Prompts und die Anpassung des Ausgabeformats auf die bei *pictory.ai* erwartete Form.

Nach weiterer Recherche wurde offensichtlich, dass dies bereits mit Erfolg praktiziert wird: insbesondere Videos die von Entdeckungen und technischen Durchbrüchen berichten sind erfolgreich [6]. Die Monetarisierung solcher Videos gestaltet sich trotz dessen schwierig.

Legislative

Die Frage, ob für KI-Systeme Regeln gelten sollen erübrigt sich. Es ist bisher nicht erlaubt einen Menschen einer algorithmischen Entscheidung ohne menschliche Prüfung zu unterwerfen. [7] Nun möchte man weiter gehen und KI-Systeme in Risikogruppen einordnen: Unannehmbares, hohes, begrenztes Risiko und Generative KI.

Gefordert wird für Generative KI, dass: [8]

1. offengelegt wird, dass ein Inhalt durch KI generiert wurde,
2. die Erzeugung illegaler Inhalte verhindert werden soll und
3. die urheberrechtlich geschützten zum Training verwendeter Daten zusammengefasst veröffentlicht werden.

Zu 1. gibt es wohl keine technisch sichere Lösung. Etwa Watermarking kann entfernt werden. Jedoch ist es möglich Interaktionen mit generativer KI für den Anwender zu kennzeichnen.

Zu 2.: Ein LLM (das auf nicht gründlichst manuell gefilterte Daten trainiert wird) kann durch Fine-Tuning nicht dazu gebracht werden illegale Inhalte nicht zu generieren. Es bedarf also mindestens weiterer Filter um das LLM. Diese können

bei Firmen wie OpenAI sicher durchgesetz werden (inwiefern dies politisch durchsetzbar ist soll hier nicht diskutiert werden), jedoch bleiben spätestens private oder kleine Akteure von diesen Regeln de facto unberührt.

Zu 3.: OpenAI verwendete zum Training öffentlich verfügbare Daten. Jedoch wurden jegliche Rechte auf intellektuelles Eigentum ignoriert. In Europa ist Text and Data Mining jedoch unter bestimmten Voraussetzungen legal, jedoch müssen Urheber die Möglichkeit besitzen der Nutzung zu widersprechen [9]. Nach aktuellem Kenntnisstand existiert keine Möglichkeit dies für generative Modelle von OpenAI zu tun.

Auditing LLMs

Ein Vorschlag [10] möchte in drei Schritten die Sicherheit von LLMs gewährleisten:

Der Technologie-Provider (bspw. OpenAI) wird hinsichtlich der Firmenstruktur, Anreize für Mitarbeiter und die Firma und Kontrollsysteme untersucht. Dabei wird ein voller Zugang zu allen Informationen im Unternehmen vorausgesetzt (white-box auditing). Dies zielt insbesondere auf die Mitarbeiter ab: wer entwickelt was mit welcher Intention. Jedoch dürfen diese Informationen nicht geteilt werden [10].

Zweitens wird das Modell überprüft. Man testet bspw. mit Red-Teaming das System auf Anfälligkeiten. Hier wird auf bspw. Jailbreaks getestet.

Drittens überprüft man die Applikation, in der ein LLM verwendet wird. Man überprüft, ob die gegebene Anwendung Nutzer diskriminiert oder ein Vorurteil besitzt und wie sie sich auf die Nutzer, Gruppen, die Gesellschaft und die Umwelt auswirkt [10] (Seite 10).

Die vorgeschlagene Methodik ist praktisch nicht realisierbar. Es fehlt primär eine Institution, die dies international und für alle Applikationen der freien Wirtschaft umsetzen könnte. Sekundär sind die Audits nicht spezifiziert: die Fragen welche Verhaltensweisen erwünscht und legal (und welche Verhaltensweisen diskriminierend) sind bleiben offen – insbesondere wie ein internationaler Konsens darüber herrschen soll.

Fazit

LLMs bedrohen die Qualität von Hilfe-Plattformen für Code (bspw. StackOverflow). Genannter Anbieter verbietet inzwischen das publizieren von ChatGPTs Antworten. Trotz vielseitiger Möglichkeiten scheint es jedoch unwahrscheinlich, dass in einem größeren Ausmaß als zuvor Manipulationen der öffentlichen Meinung in Form von gefälschten Artikeln möglich ist. Ein deutlicher Trend zeichnet sich jedoch beim Generieren von Inhalten (wie YouTube-Videos und -Shorts) ab: das Zusammenspiel verschiedener Sinne lenkt von der mangelhaften Qualität und Widersprüchlichkeit der Informationen ab. Hier ist Skepsis notwendig – jedoch fällt es schwer diese dem durchschnittlichen Konsumenten zuzuschreiben.

Bibliographie

[1] OpenAI, “GPT-4 system card.” https://cdn.openai.com/papers/gpt-4-system-card.pdf

[2] OpenAI, “Openai content policy.” https://openai.com/policies/usage-policies

[3] A. Narayanan, and S. Kapoor, “Chatgpt is a bullshit generator. But it can still be amazingly useful. https://www.aisnakeoil.com/p/chatgpt-is-a-bullshit-generator-but

[4] D. Motz, “The long road to AGI.” https://www.daniel-motz.de/articles/agi-poster

[5] Technology Innovation Institute der VAE, “Falcon-40b.” https://huggingface.co/tiiuae/falcon-40b

[6] K. Hill, “YouTube's Science Scam crisis.” https://www.youtube.com/watch?v=McM3CfDjGs0

[7] Fachgruppe Rechtsinformatik der Gesellschaft für Informatik e.V. , “Gutachten: Technische und rechtliche betrachtungen algorithmischer entscheidungsverfahren.” https://gi.de/fileadmin/GI/Allgemein/PDF/GI_Studie_Algor

[8] European Comission, “KI-Gesetz erste Regulierung der künstlichen Intelligenz.” https://www.europarl.europa.eu/news/de/headlines/

society/20230601STO93804/ki-gesetz-erste-regulierung-der-kunstlichen-intelligenz

[9] European Commission, "Intellectual property in Chatgpt." https://intellectual-property-helpdesk.ec.europa.eu/news-events/news/intellectual-property-chatgpt-2023-02-20_en

[10] Mökander, Schuett, Kirk, and Floridi, "Auditing large language models: A three-layered approach." https://arxiv.org/pdf/2302.08500.pdf