Privates Franziskus-Gymnasium
Franziskusweg 1
52393 Hürtgenwald-Vossenack

Chatbots im Ring!

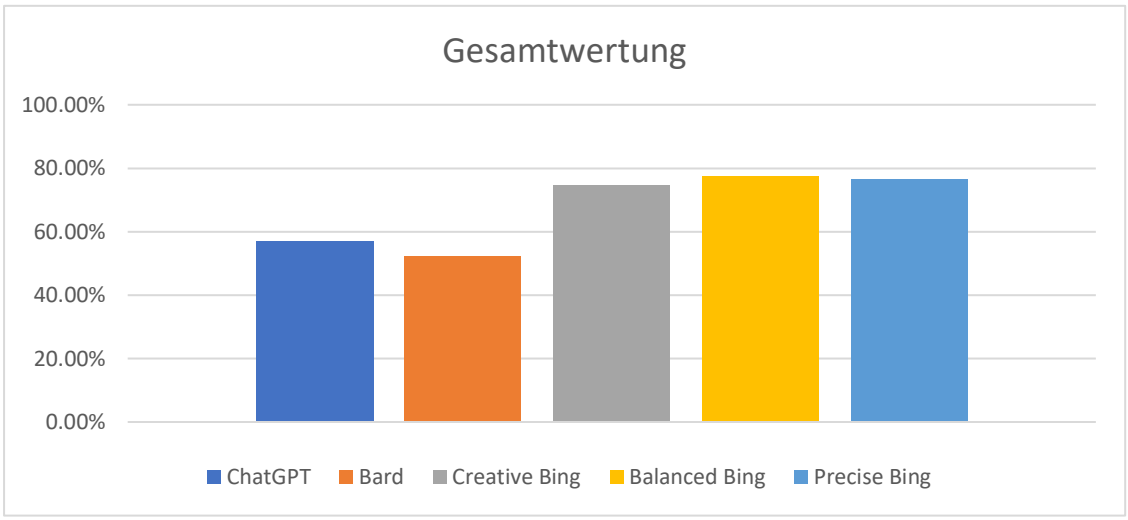
Verfasser	Luca Erkens, Miguel Falter, Simon Gillessen, Nicolas Helbig, Felix Kever, Joel Körfer, Nick Müller
Fachlehrerin	Bettina Lütten
Veröffentlichungsdatum	28. Januar 2024

Kurzübersicht

Im Folgenden wird eine systematische Untersuchung durchgeführt mit dem Ziel, die Eignung verschiedener Chatbot-Services im schulischen Kontext zu überprüfen. Bei den getesteten Services handelt es sich um OpenAI ChatGPT, Microsoft Bing Chat in drei Konversationsmodi und Google Bard. Unabhängig voneinander werden diese Chatbots in Bezug auf verschieden gewichtete Kategorien getestet, darunter unter anderem Antwortkorrektheit, Erklärfähigkeiten, Mathematikfähigkeiten und Aktualität.

Um eine möglichst objektive Wertung zu gewährleisten, wurde das Testverfahren auf Kleingruppen aufgeteilt, wobei jede Gruppe die Antworten in einem den Anforderungen entsprechenden Format dokumentiert hat. Die Ergebnisse der Dokumentation liegen vollständig vor und werden aufgrund ihrer Masse exemplarisch im anbei liegenden Video thematisiert. Somit wurden diese hauptsächlich im Textformat und bei Kategorien, welche eine Wertung der Formatierung der Antwort voraussetzten, im Bildformat, festgehalten. Alle Chatbots erhielten die gleichen Fragen und Anweisungen (Prompts) zur etwa selben Zeit, um einen Vorteil durch kontinuierliche Weiterentwicklung zu vermeiden. Zum Zugriff auf die Services wurde dabei jeweils ein Account erstellt, den alle Kleingruppen nutzten, soweit möglich, da Bard nicht auf allen Accounts zur Verfügung stand, mehr dazu im Folgenden. Die gegebenen Antworten wurden anschließend anhand eines gruppeninternen Punktesystems zunächst absolut, anschließend, basierend darauf, relativ bewertet. Dabei wurden unter Umständen einige Aspekte verschieden gewichtet aufgrund ihres variierenden Einflusses auf die Antwortqualität. Die relative Wertung ermöglichte ein abermaliges Gewichten und Aufsummieren der Ergebnisse, die jeder Service generierte, wodurch sich die Gesamtwertung ergab.

Die Recherche ergibt, dass Bing sich themenübergreifend in sehr vielen Fällen am besten eignet, während ChatGPT den konsistentesten und daher erwartbarsten Output liefert und Bard oftmals unzuverlässige Antworten gibt.
Innerhalb der Bings schneidet Creative Bing in der Gesamtwertung am schlechtesten ab, da es auf eine Fangfrage hereingefallen ist, was ein möglicher Hinweis auf eine vernachlässigte wissenschaftliche Arbeitsweise dieses Bots ist. Rechnet man die Fangfrage hinaus, ist Creative Bing allerdings wesentlich besser geeignet als die anderen zwei Bing



Modi.

Inhalt

Einleitung

Mit der Veröffentlichung des algorithmusgestützten Chatbots ChatGPT Ende November 2022 zog OpenAI die öffentliche Aufmerksamkeit über das gesamte Jahr 2023 hinweg auf das Thema der künstlichen Intelligenz. Zu Beginn noch in hohen Tönen gepriesen, begannen immer mehr Quellen, über potenzielle Probleme in der Qualität des Outputs zu berichten, was unter anderem an den zuvor etablierten hohen und teils unrealistischen Erwartungen an die unterliegenden Modelle lag.
Zeitgleich sahen viele weitere Unternehmen sich jedoch dazu veranlasst, ihre eigenen Produkte mit algorithmischer Unterstützung zu bestücken, welche dann als „KI-gestützt" vermarktet werden sollten. So auch die Google LLC und Microsoft Corporation. Erstere nutzt nun ein Sprachmodell für die gleichnamige Suchmaschine, während Microsoft sein Sprachmodell auch außerhalb der eigenen Suchmaschine, Microsoft Bing, in Form des Microsoft Copilots anbietet, welcher bald im kompletten Microsoft Ökosystem Unterstützung in Form von Zusammenfassungen und sonstiger Mitarbeit, inklusive Interaktion mit der Benutzeroberfläche, bieten soll. Im Folgenden beschränkt sich die Untersuchung jedoch auf die Chatbots der Suchmaschine. Auch hier bietet Microsofts Lösung jedoch eine Besonderheit: Der Chatbot ist in drei verschiedene „Stile" aufgeteilt, welche verschiedene Pre-Prompts, also eingängliche Instruktionen, erhalten und auf unterschiedliche Modelle setzen. Daher sind im Folgenden vertreten:

- OpenAI ChatGPT in der kostenfrei verfügbaren Version, aufgrund der Annahme, dass der durchschnittliche Schüler kein monatliches Abonnement abschließen möchte, zum Testzeitpunkt basierend auf dem Modell GPT-3.5 Turbo, ohne aktuelle Pläne, das Modell umzustellen.
- Microsoft Bing Chat in allen drei Versionen, namentlich genannt mit ihren englischen Bezeichnungen, da Microsoft regelmäßig deren Übersetzungen ändert: Creative, Balanced und Precise, im Folgenden bezeichnet mit dem Stilnamen, gefolgt von „Bing". Sowohl Creative als auch Precise nutzen aktuell das Modell GPT-4, also jenes Modell, welches OpenAI zahlenden Kunden zur Verfügung stellt, wobei im Testzeitraum das neuere Modell GPT-4 Turbo für wenige Nutzer zufällig ausgerollt wird. Unserem Testaccount wurde der Zugang nicht gestattet. Balanced nutzt „verschiedene Modelle, inklusive GPT-4"[1].
- Google Bard, welches zum Testzeitraum PaLM 2 als zugrundeliegendes Sprachmodell nutzt, aktuell sind Pläne bekannt, Gemini Pro einzusetzen. Dies wird aktuell im englischsprachigen Raum getestet, daher ebenso nicht auf unseren Accounts.

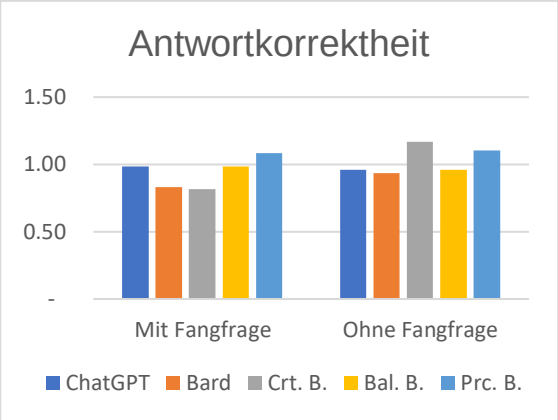
Testverfahren

Für die durchgeführten Tests wurden sowohl für ChatGPT als auch für Bing separate Accounts erstellt. Die Besonderheit im Falle von Bard bestand darin, dass Bard nicht auf dem separat erstellten Account aktiviert war, ebenso wurde der Zugriff von mehreren unserer privaten Accounts verweigert, ohne ersichtlichen Grund – keine der angegebenen möglichen Begründungen traf zu. Somit wurde Bard auf mehreren Privataccounts getestet, wobei keine der Prompts Bard aufgrund dieser Tatsache einen Vorteil oder Nachteil verschaffen sollten.
Die Antworten wurden anschließend in einem passenden Format (üblicherweise Text, sollte die visuelle Darstellung der Antworten bewertet werden, aber auch Screenshots) dokumentiert und mit Hilfe eines vordefinierten Punktesystems ausgewertet, welches anschließend normiert wurde, um einen direkten Vergleich zu ermöglichen.

[1] Mikhail Parakhin auf X, ehemals Twitter

Testergebnisse

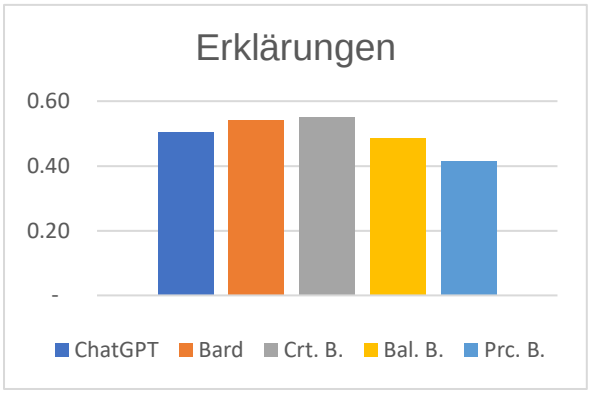
Korrektheit der Antworten



Die wohl wichtigste Eigenschaft eines Sprachmodelles, welche es zu schulischen Zwecken überhaupt erwägenswert macht, ist, dass der generierte Output faktenbasiert ist und möglichst wenig halluziniert. Halluzination bezieht sich hierbei auf die Eigenschaft von Sprachmodellen, in ihrer Funktion, Wort für Wort einen menschlich anmutenden Satz zu konstruieren, einen Output zu generieren, der grammatikalisch alle Konventionen erfüllt und durch seine Stringenz eine tatsächlich falsche Information als richtig darstellt, was durch die oben beschriebene Funktionsweise nicht immer einfach einzudämmen ist.
Um dies überprüfen zu können, wurde ein Punktesystem entwickelt, welches nach Möglichkeit die Balance trifft, Nebeninformationen zu belohnen, ohne es zu ermöglichen, eine falsche Aussage durch viele unwichtige Informationen vorteilhaft bewerten zu lassen.
Bei simplen Fragen schneiden alle Sprachmodelle weitestgehend ähnlich ab, wobei alle Modelle Fragen wie „Was passiert mit einem Fallschirmspringer, der aus einem Flugzeug springt?“ oder „Wie sieht eine gute Schweißnaht aus?“ korrekt beantworten und sich ihre Punkteanzahl dann nur dadurch unterscheidet, dass sie verschiedene zusätzliche Informationen hinzufügen.
Zu beobachten ist, dass Bard generell bei logischen Fragen am schwächsten abschneidet, während ChatGPT und Balanced Bing sich im Mittelfeld bewegen. Creative Bing punktet mit sehr ausführlichen Antworten, während Precise Bing nicht minder viele zusätzliche Informationen liefert, diese aber deutlich kompakter formuliert.
Ein ähnliches Bild ergibt sich bei doppeldeutigen Fragen: „Kann ein Känguru so hochspringen wie der Kölner Dom?“ Alle Sprachmodelle erkennen den möglichen logischen Fehler und geben eine korrekte Antwort, abgesehen von Bard. Bard hingegen gelangt an einen Punkt, an welchem es mit einstelligen Höhen des Kölner Doms in Metern rechnet und folglich auch die Frage nicht zufriedenstellen beantworten kann.
Ähnlich schwach erweist sich Bard bei unserer Fangfrage, welche keine korrekte Lösung hat: „Ein Fallschirmspringer mit 80kg springt aus 3500 Metern Höhe aus einem Flugzeug. Wie groß ist die Wellenlänge?“ Bard stellt nicht-sinnvolle Spekulationen an, wie man die Wellenlänge eines Fallschirmspringers definieren könnte, kommt aber immerhin auf die Lösung, dass die Frage nicht sinnvoll zu beantworten sei. Problematischer wird aber die Antwort von Creative Bing, welches nicht erkennt, dass die Frage keine sinnvolle Antwort bietet und erfolglos versucht, glaubwürdig eine Wellenlänge zu errechnen, wobei es, dadurch, dass keine vernünftige Antwort möglich ist, auf dem Weg mehrere Fehler macht.
Beide Chatbots bekommen in der Kategorie „Fachwissen“, in welchen Bard und Creative Bing als Sieger hervorgehen, durch die genannten Fehler, Abzüge, während Balanced Bing und Precise Bing eine leicht schlechtere Leistung erbringen und ChatGPT einen Teil der Frage ignorierte, somit weniger Punkte erhält.
Zusammenfassend bietet Precise Bing in der Hinsicht der Korrektheit die besten Antworten, gefolgt mit einem moderaten Abstand von ChatGPT und Balanced Bing, dann Bard und anschließend Creative Bing. Wird die Fangfrage nicht berücksichtigt,

ändert sich daran nicht viel, abgesehen davon, dass Creative Bing nicht das Schlusslicht bildet, sondern den ersten Platz belegt. Es ist anzumerken, dass, während Balanced und Precise Bing häufig die konsistentesten Antworten liefern, auch, wenn sie nicht immer die höchstbepunkteten sind. Sie sind häufig dennoch die verlässlichsten Chatbots.

Erklärung von komplexen Thematiken



Durch den schulischen Kontext ist eine korrekte Antwort zwar sehr wichtig, aber nur eingeschränkt nützlich, sollte der Chatbot nicht in der Lage sein, die besprochenen Konzepte nachhaltig zu erklären.
Die Sprachmodelle bekommen jeweils dieselbe Frage gestellt. Diese Frage muss, auf Nachfrage, jeweils auf Unter-, Ober- und Mittelstufen- sowie Universitätsniveau beantwortet werden. Wie bereits oben werden die Ergebnisse anschließend in einer Tabelle festgehalten, um eine Aussage darüber treffen zu können, welches Sprachmodell sich für welche Bildungsstufe am besten eignet.

Thematisches Niveau

Alle fünf Bots haben Probleme, sich sowohl dem Universitäts-, als auch dem Unterstufenniveau anzupassen. Zur gleichen Zeit funktioniert die Anpassung an das Mittel- und Oberstufenniveau hingegen gut. Creative Bing geht in dieser Kategorie als bestbewertet hervor, gefolgt von ChatGPT. Balanced Bing bildet zusammen mit Bard das Schlusslicht.

Einbringen praktischer Beispiele

Sofern die Modelle Beispiele einbinden, um Sachverhalte verständlicher zu erläutern, sind diese gut gelungen, allerdings geschieht dies auf Universitätsniveau kaum, die meisten Beispiele werden auf Unter- und Mittelstufenniveau verwendet, was verständlich ist. In dieser Kategorie belegt erneut Creative Bing den ersten Platz, gefolgt von ChatGPT. Bard befindet sich auf dem letzten Platz, da es keine Beispiele verwendet.

Einbinden von Materialien

Die Einbindung von visuellen Materialien in die Erklärung geschieht bei allen Modellen auf enttäuschende Weise. Abgesehen von Bard stellt kein Modell eine solche Funktion bereit, wobei Bard, dadurch, dass es die Bilder nicht selbst sieht, auch auf Unterstufenniveau eine große Menge komplexer Bilder hintereinander in seine Antwort einbettet, die allesamt ähnliche Inhalte zeigen. Daher erhält Bard eine mäßig gute Wertung, alle anderen Modelle eine schlechte.

Fachsprache

Fachsprache ist ein wesentlicher Bestandteil jedes Faches, auf einem höheren Niveau wird diese unabdinglich, dementsprechend macht es Sinn, dass, auch wenn die Sprachmodelle ihre Verwendung von Fachsprache sehr unterschiedlich handhaben, grundsätzlich zu erkennen ist, dass mehr als die Hälfte von ihnen auf Oberstufenniveau lieber darauf zurückgreift als auf Unterstufenniveau. Hier schneidet Creative Bing im Test am besten, ab, gefolgt von Bard, ChatGPT und Precise Bing, dann Balanced Bing.

Einfache Sprache

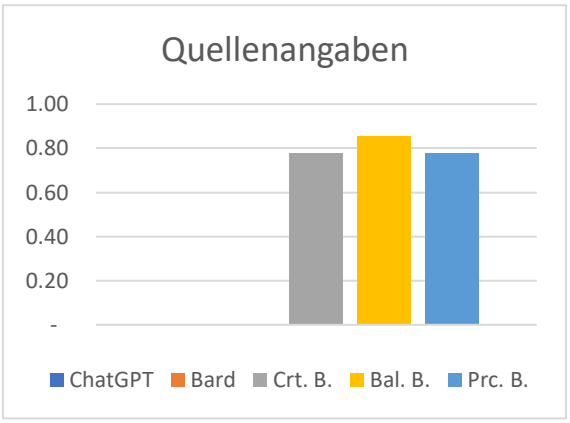
Gerade in der Unterstufe ist nicht nur Fachsprache wichtig, sondern gerade einfache Sprache. Hier erzielen alle Bots gute Ergebnisse, auch wenn die Sprache in der Unterstufe durchaus noch etwas einfacher sein dürfte. Balanced und Creative Bing erhalten hier eine gute Bewertung, Bard jedoch eine schlechte.

Struktur

Strukturell produzieren alle Kandidaten äußerst gute Ergebnisse. ChatGPT, Bard und Balanced Bing erhalten hier die volle Punktzahl, gefolgt von Creative und Precise Bing, wobei deren Struktur an sich nicht schlecht, aber dennoch auch nicht perfekt ist.

Somit lässt sich zusammenfassend sagen, dass für die Unterstufe ChatGPT und Balanced Bing am besten geeignet sind, für die Mittelstufe ebenso Balanced Bing, für die Oberstufe Bard und für die Universität Creative Bing.
Im abschließenden Vergleich der Modelle in der Hinsicht darauf, welches Modell sich am besten verhält, wenn man sich für seine gesamte Laufbahn auf eines festlegen möchte, liegt Creative Bing auf dem ersten Platz, dicht gefolgt von Bard, mit einigem Abstand folgen ChatGPT und Balanced, auch dicht zusammen, und mit einigem weiteren Abstand folgt Precise Bing, hauptsächlich dadurch, dass seine Erklärungen für komplexe Zusammenhänge zumeist auf einem derart hohen Niveau sind, dass es nicht zu rechtfertigen wäre, es zu verwenden, anstatt den Wikipedia Artikel zum Thema zu lesen, welcher vom Sprachniveau nicht weit entfernt ist.

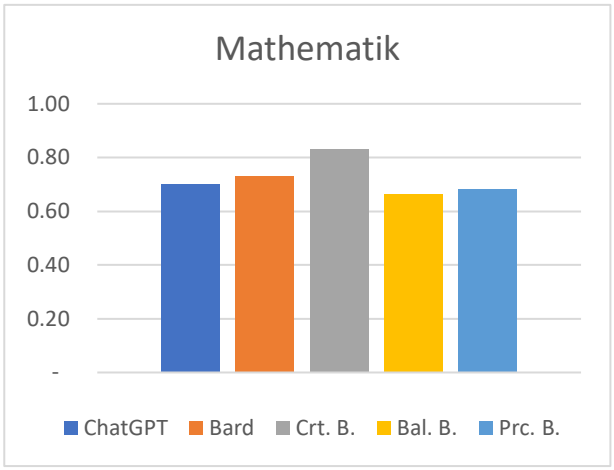
Quellenangaben



Auch außerhalb von wissenschaftlichen Arbeiten ist es wichtig, dass Sprachmodelle ihre Quellen verlässlich angeben können, zum einen, um ihre Aussagen zu belegen und zum anderen, damit man Quellen, welche sie verwenden, selbst verwenden kann.
Die Tatsache, dass ChatGPT standardmäßig keine Quellen angibt und deren Angabe sogar verweigert, ist bedingt dadurch, dass es sich um ein Sprachmodell ohne Zugang zu aktuellen Quellen handelt. Viel überraschender ist die Tatsache, dass Bard sich ähnlich verhält. Weder gibt es standardmäßig Quellen an, noch tut es dies bei expliziter Nachfrage. Dabei ist anzumerken, dass Bard grundsätzlich in der Lage ist, Quellen anzugeben, diese Funktion ist aber absolut nicht gut ausgearbeitet und die Quellen, die Bard innerhalb anderer Tests angegeben hat, passten selten auch nur entfernt zu seinen Aussagen.
Auf der anderen Seite gibt Bing standardmäßig Quellen an, wodurch keine Nachfragen nötig sind. Dabei ist zu beachten, dass Precise Bing stets versucht, sich so gut wie möglich auf etablierte Quellen zu beziehen, etwa das Bundesamt für politische Bildung oder The Guardian. Balanced Bing versucht dies ebenso, greift jedoch zusätzlich auch auf weniger etablierte Quellen zurück, um mehr Informationen zu erhalten. Bei Creative Bing ist die Wahrscheinlichkeit, dass es weniger etablierte Quellen nutzt, am höchsten. Zu einer Frage hat es etwa nur unbekannte Seiten genutzt, aber dennoch die richtige Antwort gegeben. Balanced Bing gibt zudem einmal eine Quelle an, ohne, dass die Information, die es angibt, aus dieser Quelle stammt.

Dadurch, dass Balanced Bing allerdings die meisten Quellen angibt, liegt es in diesem Test vorne, dahinter sind Precise Bing und Creative Bing gleichauf. Sowohl Bard als auch ChatGPT erhalten hier null Prozent.

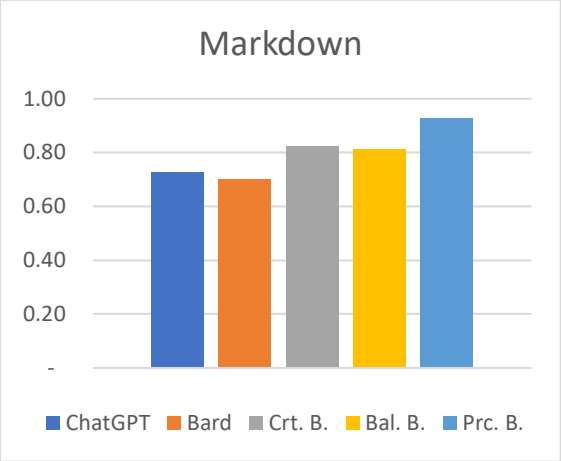
Mathematische Fähigkeiten



Dadurch, dass Sprachmodelle funktionieren, indem sie das nächste Wort erraten, sind sie in der Lage, rudimentär mathematische Probleme zu lösen. Mit steigender Komplexität sinkt die Wahrscheinlichkeit, dass die korrekte Lösung erraten wird, jedoch erwartungsgemäß.
Bei ebendiesen simplen Aufgaben ist Bard in der Lage, die besten Ergebnisse zu liefern, jedoch nur mit einem vernachlässigbaren Vorsprung. Bei höheren Graden der Komplexität produzieren die Sprachmodelle kein einziges richtiges Ergebnis mehr, trotz genauer Beschreibung der Vorgehensweise. Kein Ergebnis anzugeben, und, wie in Precise Bings Fall, sogar darauf hinzuweisen, dass nun ein Taschenrechner von Nöten sein wird, um das Endergebnis zu berechnen, spricht jedoch für das Modell, ihm seine Limitationen zu einem gewissen Grad bekannt zu sein scheinen, da es das Modell erfolgreich davon abhält, ein falsches Ergebnis zu halluzinieren.
Zusammenfassend ist allerdings zu sagen, dass keiner der Chatbots für höhere Mathematik tatsächlich gut geeignet ist, da sie alle an unterschiedlichen Stellen Fehler machen, wodurch am Ende kein verlässliches Ergebnis bei der Rechnung herauskommt. Auch wenn einfachere Fragestellungen mit einer hohen Wahrscheinlichkeit richtig beantwortet werden, ist es ratsam, jeden Rechenschritt noch einmal zu überprüfen. In der Mathematik und Physik bietet sich also eine kollaborative Arbeitsweise mit den Modellen an, anstelle einer solchen, die die Chatbots mit einer Aufgabe allein lässt und sie mit der Aufgabe betraut, diese ohne fremde Hilfe zu lösen.
Es ist anzumerken, dass ChatGPT+ aktuell ein Feature besitzt, welches ChatGPT Code schreiben lässt, welcher dann von Python Interpreter ausgeführt wird, wodurch Berechnungen wesentlich verlässlicher funktionieren sollten, da sie tatsächlich berechnet und nicht erraten werden. Microsoft plant, dieses Feature ebenfalls in Form eines Python Code Interpreters im Jahr 2024 auch in Bing Chat, beziehungsweise Microsoft Copilot, einzubauen, weshalb Hoffnung auf Besserung besteht[2]. Momentan ist nicht bekannt, dass Google und OpenAI planen, dieses Feature in näherer Zukunft ebenso in die kostenfreien Versionen ihrer Produkte einzupflegen.

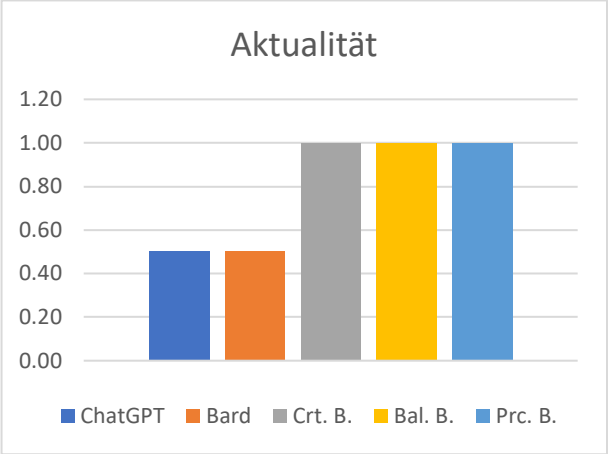
[2] Mehdi 2023: Celebrating the first year of Copilot with significant new innovations.

Markdown



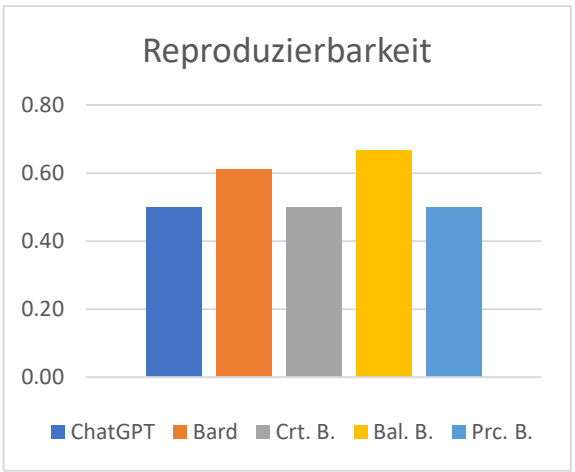
Markdown-Fähigkeiten sind ein wichtiges Element in der Präsentation von Antworten. Sei es das Fettdrucken von wichtigen Informationen, um eine schnelle Beantwortung der Frage, ohne die gesamte Antwort zu lesen, zu ermöglichen oder seien es Tabellen, welche eine Übersicht über Größen, Einheiten und Beschreibungen für den Physikunterricht oder eine Auflistung an Stilmitteln für den Deutschunterricht bereitstellen.
Hier ist anzumerken, dass Bing grundsätzlich die meisten Features bietet. Abgesehen von Bildeinbettungen in Antworten fehlt Bing kein Feature, welches ein anderer Chatbotservice bietet. Hinzu kommt, dass Bing dieses Feature in der Vergangenheit besaß, ebenso wie eine Google Charts-Integration, um Graphen zu erstellen – beide wurden deaktiviert.
Die Tatsache, dass Bing laut dem eigenen Pre-Prompt (Instruktionen, die der Chatbot erhält, die ihm erklären, wie es sich zu verhalten gilt, daher auch als System Prompt, Regeln oder Instruktionen bezeichnet) „GitHub Flavored Markdown“ verwendet, wird Balanced Bing zum Verhängnis, da es versucht, Bilder einzubetten, dann jedoch in einer Endlosschleife gefangen wird, in welcher es versucht, einen unendlich langen Link zu tippen, der sowieso vom System nicht erkannt werden könnte.
Gleichzeitig ist anzumerken, dass Bing versucht, seine Antworten sehr schön zu formatieren mit einem Bild des Subjektes in der oberen rechten Ecke oder weiteren Bildern in einem Widget unter der Antwort. Auch wenn dies keine Funktion ist, die das Sprachmodell selbst steuert, ist es dennoch von Ästhetik geprägt.
Bard geht davon aus, sehr viele Features zu haben, die es nicht hat. Wenn es versucht, diese zu benutzen, funktioniert dies entsprechend nicht. Beispielsweise versucht Bard sehr oft, mathematische Ausdrücke mithilfe von LaTeX zu formatieren, allerdings bettet es sie stattdessen in Codeblöcke ein, wodurch sie nicht richtig angezeigt werden. Dasselbe gilt für andere Features.
ChatGPT verwendet Markdown-Fähigkeiten sehr verlässlich, Bing vergisst manchmal, dass sie existieren, oder macht kleine Syntaxfehler, während Bard hier recht inkonsistent ist.
Insgesamt, allein aufgrund der Anzahl der Features, erhält Precise Bing die höchste Wertung, gefolgt von Creative und Balanced Bing, beide recht nah beieinander mit einem Abstand und dann, nach einem weiteren Abstand, folgen ChatGPT und zuletzt Bard. Die Gesamtwertung ist durch Evaluieren der jeweiligen Nützlichkeit der einzelnen Tools zu errechnen. So ist die Fähigkeit, LaTeX-Ausdrücke zu verwenden, deutlich hilfreicher als die Nutzung von Emojis, daher wird sie auch höher bewertet.

Aktualität



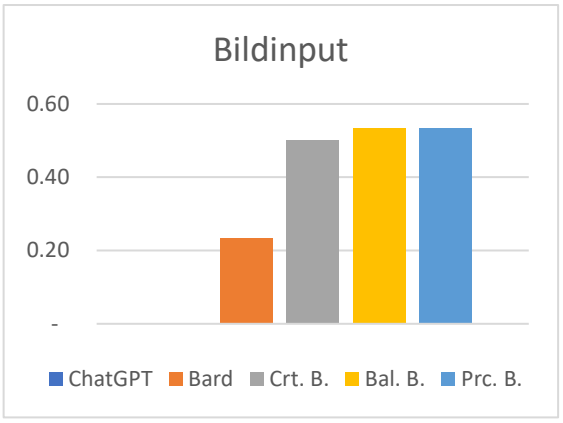
Insbesondere bei der Recherche zu aktuellen Themen oder Sachverhalten, die sich ohne Kenntnis des Nutzers geändert haben, ist es hilfreich, wenn der Chatbot in der Lage ist, auf aktuelle Informationen zurückzugreifen.
Zu diesem Zweck besitzen Bard und Bing eine Funktion, mit der sie mehr oder minder auf Echtzeitinformationen zurückgreifen können, mit der Einschränkung, dass beide, aus Sicherheitsgründen und um eine hohe Antwortgeschwindigkeit zu gewährleisten, den entsprechenden Webseitencaches der Suchmaschine lesen und nicht tatsächlich die Webseite besuchen, anders als die Browserfunktion in ChatGPT+, welche sich unter anderem daher als sehr langsam erweist.
ChatGPT und Bing sind in der Lage anzugeben, bis wann ihre internen Daten aktuell sind, wobei der interne Datenstand für Bing mittlerweile aktualisiert ist, das hinterlegte Datum im Pre-Prompt allerdings nicht, womit insbesondere Precise Bing, welches dazu neigt, sein „Cutoff-Date“ öfter als nötig zu erwähnen, da es eine recht „rohe“ Version des Modells von OpenAI darstellt, welches, wie auch die anderen Bings, mit dem Prometheus-Modell in Verbindung gebracht wird, um die Websuche zu ermöglichen.
Bard erwähnt, dass es auf einem Datensatz trainiert ist, welcher ständig aktualisiert werde, um anschließend bei Nachfrage nach einem aktuellen Thema, auf die Google Suchmaschine zu verweisen, anstatt selbst eine Antwort zu geben, was allerdings ein Einzelfall war. ChatGPT ist das besprochene neue Ereignis nicht bekannt und es bietet mehrere alternative Ereignisse an, welche gemeint sein könnten, Bing beantwortet die Frage korrekt.

Reproduzierbarkeit



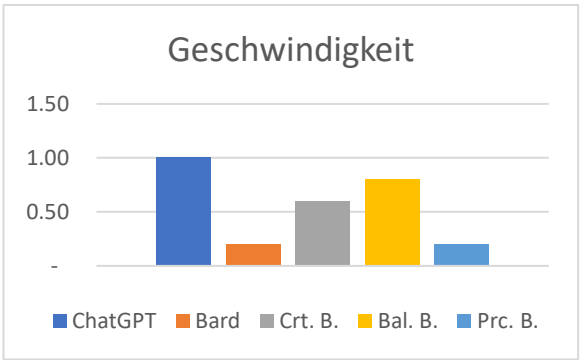
Eine nicht minder wichtige Eigenschaft eines Sprachmodelles ist, dass beim wiederholten Stellen einer Frage nicht bei jedem Versuch ein verschiedenes Ergebnis herauskommt.
Hierbei sind alle Chatbots überwiegend zuverlässig, wobei Balanced Bing den ersten Platz belegt, gefolgt von ChatGPT, seinerseits gefolgt von Bard und den verbleibenden Bings, beispielsweise gibt Bard zwei verschiedene Autos als das weltschnellste Auto an, allerdings sind die Ergebnisse ansonsten, abgesehen vom Wortlaut, immer recht ähnlich, wodurch keines der Ergebnisse an sich schlecht ist – auf einer Skala von 0% bis 100% liegen alle innerhalb einer Spanne von deutlich unter 20 Prozentpunkten.

Bildinput



Ein möglicherweise nützliches Feature ist es, der KI ein Bild zu schicken und sich darauf basierend mithilfe des multimodalen Inputs des Services eine Erklärung liefern zu lassen. ChatGPT fehlt dieses Feature gänzlich.
Hier ist es erstaunlich, wie weit Bard oftmals von der tatsächlichen Lösung entfernt ist. Am erstaunlichsten ist, dass es bei einem Bild einer Katze, auf der eine Orange liegt, was insgesamt so aussehen soll, wie ein Spiegelei, korrekt erkennt, dass es sich um eine Katze handelt, dann aber sagt, dass ein Spiegelei auf der Katze läge. Unsere Annahme besteht darin, dass es online nach dem Bild sucht, ein Feature, welches Bing fehlt, aber noch nachgeliefert werden soll[3], und dann durch eine Beschreibung, die ein Spiegelei enthielt, durcheinandergerät. Ebenso glaubt Bard, bei dem Bild einer Würfelqualle die Quallenart erkannt zu haben, beschreibt dann aber das Aussehen einer Qualle, welche laut Bards Beschreibung eine komplett andere Farbe als jene im Bild hat. Auf Nachfrage wiederholt sich das Phänomen. Während die Bings bei der ersten Frage noch eine korrekte Antwort liefern, erzielen diese hier keine Punkte, obgleich ihre Antworten, statt komplett danebenzuliegen, plausibler, aber dennoch falsch oder zumindest ausweichend sind, auch wenn letzteres nicht gegen den Chatbot spricht. Auch die Fähigkeit, ein handgeschriebenes lineares Gleichungssystem zu erkennen und anschließend zu lösen oder eine Strukturskizze zu erklären, hält sich bei allen sehr stark in Grenzen.

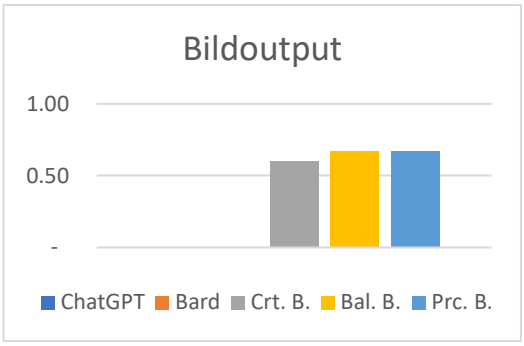
Geschwindigkeit



Geschwindigkeit spielt eine eher untergeordnete Rolle im Kontext von KI-Assistenten, allerdings kann sie dennoch ab und an wichtig sein, wenn nicht viel Zeit für die Beantwortung einer Frage zur Verfügung steht. Hierbei wird nicht gemessen, wann eine Antwort fertig geschrieben ist oder wann das Modell beginnt, sie zu schreiben, sondern wann die tatsächliche Antwort auf dem Bildschirm erscheint.
ChatGPT gibt auf die ihm gestellte Testfrage zwar die falsche Antwort, ist aber mit Abstand am schnellsten. Bard braucht etwas länger und alle Bings sind recht langsam, wobei bei Bing die Auslastung der Server am meisten zu spüren ist und daher die Zeit am inkonsistentesten ist. Weiterhin soll Bing aber mit GPT-4 Turbo ausgestattet werden, was dieses Problem zumindest eindämmen sollte, aus mehreren Gründen: Einmal, da das Modell schneller ist und außerdem, da es effizienter ist.

[3] Ebd.

Bildoutput

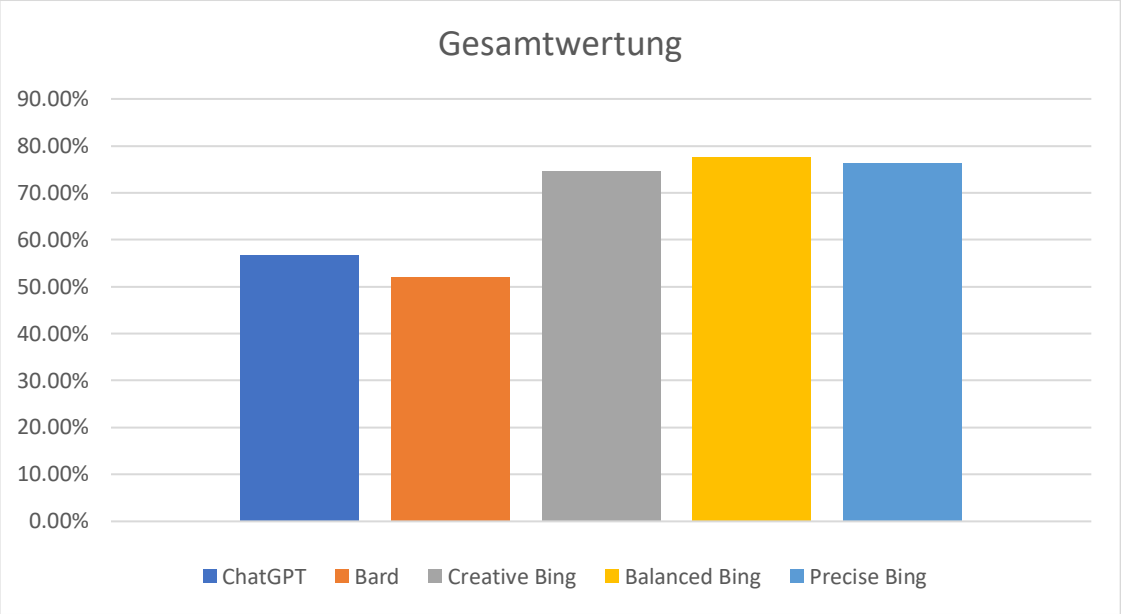


Diese Kategorie trägt einen noch geringeren Wert, allerdings soll sie der Vollständigkeit halber nicht unerwähnt bleiben.
Bing ist der einzige getestete Service, welcher in der Lage ist, Bilder zu generieren. Dies geschieht mithilfe von OpenAIs Dalle-3, obgleich manche Stimmen in der Community glauben, dass eine abgespeckte Version zum Einsatz kommt. Anzumerken ist, dass der Bing Image Creator ein sehr radikales Filtersystem besitzt, standardmäßig werden vier Bilder generiert, allerdings wird alles geblockt, was der Filter als Verstoß gegen die Inhaltsrichtlinien einstuft. Gleichzeitig ist es aber augenscheinlich dennoch sehr einfach, solche Bilder zu generieren, wodurch die Erfahrung für alle Nutzer suboptimal ist, die Bilder generieren möchten. Zudem ist der Filter nicht besonders adaptiv und blockt basierend auf Wörtern. Im Subreddit häufen sich ebenso die Beschwerden, dass Prompts mit Frauen geblockt werden, dann aber ohne Änderungen durchgehen, ohne, dass irgendeine Änderung an ihnen stattfindet, abgesehen davon, dass alle Charaktere männlich gemacht werden. Das hängt mutmaßlich sowohl mit dem Prompt Blocker zusammen als auch mit den Trainingsdaten des verwendeten Modells, wodurch das Modell scheinbar sehr schnell dazu gebracht wird, unangebrachte Inhalte zu generieren, die das System dann blockiert. Dennoch ist das Feature, Bilder zu generieren eine nette Dreingabe, die zwar nicht wirklich hilfreich, aber auch nicht schlecht ist. Sie funktioniert exakt gleich zwischen allen drei Bings, weshalb alle Outputs ungefähr gleich gut sind, auch wenn komplexe Prompts mit mehreren Charakteren, die miteinander interagieren, nicht immer richtig erkannt werden.

Fazit

Für eine abschließende Bewertung sind die normalisierten Einzelwertungen zu summieren, um einen finalen Durchschnittswert zu errechnen. Zu diesem Zwecke wird die Antwortkorrektheit zehnfach gewichtet, Quellenangaben, Pädagogik und Mathematik siebenfach, Markdown-Fähigkeiten sechsfach, Aktualität fünffach, Reproduzierbarkeit und Bildinput dreifach, Geschwindigkeit zweifach und Bildoutput einfach. Diese Faktoren sind abgeleitet von ihrer jeweiligen Nützlichkeit im schulischen Kontext.
Zum Ende erreicht ChatGPT eine Wertung von 57%, Bard 52%. Bings Ergebnis fällt deutlich positiver aus mit 75% für Creative Bing, 76% für Precise Bing und 77% für Balanced Bing. Dabei ist anzumerken, dass durch die hohe Gewichtung der Korrektheit der Antworten Creative Bings Versagen bei der Fangfrage ein sehr hohes Gewicht trägt. Berücksichtigte man die Fangfrage nicht, erhielte Creative Bing daher eine Wertung, welche sich sehr stark von jener mit Fangfrage unterscheidet, anders als bei den anderen Bots. Diese Wertung läge bei 81%.
Dieses Ergebnis mag zuerst unerwartet klingen durch ChatGPTs Popularität und Bards zahlreicher negativer Berichterstattung. tatsächlich macht Bard oftmals offensichtlichere Fehler, was allerdings nicht bedeutet, dass es zwangsläufig mehr Fehler als ChatGPT macht. Dabei ist ebenso anzumerken, dass ChatGPT die konsistentere Erfahrung bietet, es ist also in dem Sinne verlässlich, dass man sich darauf verlassen kann, einen Output zu erhalten, der die Qualität des letzten Outputs recht exakt einfängt.
Bing erweist sich in den meisten Fällen weitaus solider als ChatGPT und Bard, ihm fehlt jedoch an manchen Ecken der Feinschliff, gerade was die Benutzererfahrung angeht über jegliche Benutzeroberflächen, die Microsoft für den Chatbot anbietet. Was Bing

fehlt, ist die flüssige Benutzererfahrung, die einen komfortablen Einstieg ermöglicht. Vorwissen bezüglich der Funktionsweise von Sprachmodellen ist bei Bing nicht selten hilfreich, um interessantes Verhalten erklären zu können. Gleichzeitig bietet Bing aber auch in vielen Fällen den qualitativ wertvollsten und stimmigsten Output und die Tatsache, dass es mehr oder minder einen Charakter besitzt, wurde während unserer Tests oftmals als sehr angenehm empfunden, wodurch die Interaktion mit Bing sich oft weniger wie Arbeit anfühlt, sondern, trotz einiger kleinerer Probleme, tatsächlich Spaß macht, was bei Bard und ChatGPT in ersterem Falle unfreiwillig und in letzterem Falle



sehr selten vorkommt.

Quellenverzeichnis

1 Parakhin, M. [MParakhin] „Creative and precise a 100% GPT-4, Balanced is a combination of several models including GPT-4“ [https://twitter.com/MParakhin/status/1693579775590224097] [Tweet]
2 Mehdi, Y. „Celebrating the first year of Copilot with significant new innovations” [https://blogs.microsoft.com/blog/2023/12/05/celebrating-the-first-year-of-copilot-with-significant-new-innovations/] [zugegriffen am 18. Januar 2023]

Eigenständigkeitserklärung

Hiermit versichern wir, Luca Erkens, Miguel Falter, Simon Gillessen, Nicolas Helbig, Felix Kever, Joel Körfer, Nick Müller, dass wir den vorliegenden Testbericht selbstständig angefertigt haben. An den Stellen, an denen wir beim Testen Unterstützung oder Hilfe durch andere Personen erhalten haben, haben wir das angegeben. Hilfsmittel, benutzte Literatur oder Quellen aus dem Internet haben wir angegeben oder als Zitat kenntlich gemacht.