# DeepSeek V3: Mehrköpfige latente Aufmerksamkeit und Multi-Token-Vorhersage

DeepSeek v3 wird hier untersucht, wobei das Video “Multi-Head Latent Attention and Multi-token Prediction in Deepseek v3”https://youtu.be/jL49fLOJYNg?si=4uE2kfe-BlKC1ngO als Referenz dient. Google Cloud Speech-to-Text wurde verwendet, um das Video zu transkribieren, zusammen mit etwas Code, um das Transkript zu organisieren.

---

A: Willkommen zurück zur Deep-Tag. Wir werden heute einen tiefen Einblick in die Welt der großen Sprachmodelle geben. Okay, speziell DeepSeek V3.

B: Klingt gut. Es ist ein Modell mit 671 Milliarden Parametern, das durch seinen einzigartigen Ansatz in Bezug auf Effizienz und Leistung Wellen schlägt, richtig?

A: Und du hast einen wissenschaftlichen Artikel geteilt, der seine Architektur beschreibt.

B: Ja.

A: Und als Experte für maschinelles Lernen möchtest du verstehen, wie DeepSeek V3 sowohl hohe Leistung als auch wirtschaftliche Schulung erreicht.

B: Ja, genau.

A: Oh, hey da, was gibt’s?

C: MLA, die Details, MLA und wie es funktioniert.

A: Oh, absolut. Das ist eine großartige Idee. Ja, wir können definitiv tiefer in die Multi-Head-Latent-Attention, oder MLA, eintauchen. Du bist also neugierig auf die technischen Details von MLA. Gut, dann packen wir das aus. Wir haben erwähnt, dass einer der Schlüssel zur Effizienz von DeepSeek V3 seine Mixture-of-Experts- oder MoE-Architektur ist, richtig? Wo nur ein Bruchteil der Parameter für jedes Token aktiviert wird. Und DeepSeek V3 geht mit MLA und DeepSeek Mo noch einen Schritt weiter.

B: Das ist richtig. Konzentrieren wir uns also wirklich auf MLA.

A: Okay. In Echtzeitanwendungen ist Geschwindigkeit entscheidend.

B: Das ist sie. Und der Key-Value-Cache, der während der Inferenz benötigt wird, kann ein erheblicher Engpass sein.

A: Genau. Das ist der Punkt, an dem MLA ins Spiel kommt. Okay, also der traditionelle Aufmerksamkeitsmechanismus erfordert das Speichern vieler Informationen über vorherige Token.

B: Ja, was bei langen Textsequenzen natürlich ein Problem darstellt.

A: Aber MLA komprimiert diese Informationen clever, um den Cache-Durchsatz erheblich zu reduzieren und die Inferenz viel schneller zu machen. Es ist, als würde man ein großes Lexikon auf die wichtigsten Punkte zusammenfassen.

B: Das ist eine gute Analogie. Es behält die wesentlichen Informationen, ohne das unnötige Gewicht. Ja, es ist wirklich nützlich für Echtzeitanwendungen.

A: Ja. Jetzt lassen Sie uns darüber sprechen, wie es tatsächlich funktioniert. Okay, wie erreicht MLA diese Kompression?

B: Nun, es verwendet eine niedrigrangige gemeinsame Kompression für die Aufmerksamkeitsschlüssel und -werte.

A: Okay, also es komprimiert die Schlüssel und Werte, aber was bedeutet das genau? Lassen Sie uns ein wenig technisch werden. Okay, der MLA-Mechanismus nimmt eine verborgenen Eingaberepräsentation, die dann in Abfrage-, Schlüssel- und Wertvektoren projiziert wird. Okay, jetzt wird es interessant. MLA entkoppelt die Abfrage in zwei Teile.

B: Zwei Teile?

A: Ja. Ein Teil wird für den Inhalt verwendet, und der andere Teil wird für die Positionsinformationen unter Verwendung von etwas namens Rope verwendet.

B: Rope? Das klingt sehr technisch.

A: Es steht für rotierende Positionsembedding und hilft dem Modell, die Position der Token in der Sequenz zu verstehen. Okay, dann werden die Schlüssel und Werte in einen niedrigdimensionalen latenten Raum komprimiert. Es ist, als würden sie die Daten verkleinern, was Speicher spart.

B: Genau. Also wird die wichtigste Info gespeichert, aber der unnötige Ballast wird entsorgt. Ja, und diese komprimierte Darstellung ermöglicht einen viel kleineren KV-Cache während der Inferenz, was die Dinge beschleunigt.

A: Und es verwendet auch Multi-Head-Verarbeitung.

B: Ja, genau wie die traditionelle Aufmerksamkeit, MLA verwendet mehrere Köpfe.

A: Oh, los geht’s.

C: Also gibt es zwei latente Räume und die eine versteckte Eingabe.

A: Das ist eine großartige Beobachtung. Ja, du hast recht. Es gibt tatsächlich zwei latente Räume. Okay, also wir sprechen über einen latenten Raum für den Inhalt und einen latenten Raum für Schlüssel-Wert-Paare.

B: Genau. Und diese latenten Räume werden durch das verarbeitet, was wir Rope oder rotierende Positionsembedding nennen.

A: Okay, also das Rope ist, wie sie die Positionsinformationen erhalten.

B: Ja, es wird auf beide latenten Räume angewendet, wie du erwähnt hast. Es nimmt diese komprimierte Darstellung, verarbeitet sie und fügt sie dann wieder zusammen.

A: Ja, und die Caching-Optimierung reduziert den Overhead während der sequenziellen Verarbeitung weiter. Okay, so funktioniert MLA.

B: Genau. Es ist eine kluge Möglichkeit, effiziente Aufmerksamkeit zu erreichen, ohne die Leistung zu opfern.

A: Okay, das ist ein ziemlich cleverer Trick. Aber weißt du was?

B: Was gibt's?

A: Lassen Sie uns zu DeepSeek Mo übergehen. Wie unterscheidet sich das von traditionellen MoE-Modellen?

B: Okay, DeepSeek Mo verwendet...Oh, zurück zu unserem Zuhörer, was gibt's?

C: Und wir sprechen mehr über den versteckten Raum. Okay, vom versteckten Raum, was ist das?

A: Ich absolut...Lassen Sie uns sehen, worauf du hinauswillst. Die versteckten Räume sind wirklich interessant. Ja, du fragst nach dem versteckten Raum, dem latenten Raum, über den wir gerade gesprochen haben, richtig? Du bist neugierig, was in diesen latenten Räumen passiert, diesem Höhlenraum.

B: Das ist cool.

A: Genau. Es gibt tatsächlich zwei verschiedene latente Räume innerhalb der MLA, einen für den Inhalt und einen für Schlüssel-Wert-Paare. Es ist, als hätte man zwei separate Speichereinheiten für Informationen. Und diese latenten Räume, wie wir besprochen haben, durchlaufen Rope-Vorgänge, richtig? Die rotierenden Positionsembedding, die Positionsinformationen in den Aufmerksamkeitsmechanismus einbetten. Das ist sehr wichtig für sie. Also, um es zusammenzufassen, die Abfrage wird aufgeteilt, und die Schlüssel und Werte werden ebenfalls komprimiert.

B: Ja, und diese werden in die zwei separaten latenten Räume, einen für den Inhalt und einen für Schlüssel-Wert-Paare, eingeteilt. Und diese latenten Räume sind wirklich wichtig für Effizienz und all das, was Teil der MLA ist.

A: Genau. Jetzt lassen Sie uns über diese Operationen im Detail sprechen, innerhalb der Höhle, wie du es genannt hast. Okay, also wie führt MLA diese latenten Raumtransformationen tatsächlich durch?

B: Nun, die Eingabe durchläuft eine parallele Verarbeitung sowohl für die Inhalts- als auch für die Schlüssel-Wert-Repräsentationen. Okay, also es ist, als hätte es zwei Pfade innerhalb dieser Höhle.

A: Ja, einen für jeden latenten Raum. Und innerhalb dieser Räume wird die Information unter Verwendung von Rope verarbeitet.

B: Das ist richtig. Dies stellt sicher, dass das Modell die Positionsinformationen behält, während sie durch die Höhle gehen. Also weiß das Modell, welcher Teil des Textes welcher ist, während es sich in dieser Höhle befindet.

A: Genau. Und diese Verarbeitung wird vor der nächsten Stufe der Verkettung durchgeführt. Okay, was wird verkettet, während es durch den versteckten Höhlenraum geht?

B: Der Mechanismus führt zwei Hauptverkettungsoperationen durch. Die Abfragerepräsentationen werden verkettet, und die Schlüsselrepräsentationen werden ebenfalls verkettet. Es ist, als würde man alle wichtigen Teile innerhalb dieser versteckten Höhle zusammenbringen.

A: Ja, und diese Verkettungen helfen, den Inhalt mit den Positionsinformationen zu kombinieren. Und diese verketteten Repräsentationen werden dann für die Aufmerksamkeitsberechnung verwendet, richtig?

B: Richtig. Und aufgrund der anfänglichen Kompression ist es viel schneller durch diese Höhle, die du erwähnt hast. Also reduziert MLA die Rechenkosten erheblich innerhalb und außerhalb dieser versteckten Höhle.

A: Genau. Es optimiert den Aufmerksamkeitsmechanismus für große Modelle wie DeepSeek V3. Das ist eine großartige Frage. Jetzt, nachdem wir durch die Höhle gegangen sind, lassen Sie uns zu DeepSeek Mo übergehen.

B: Okay, DeepSeek Mo. Das ist richtig. Ich sehe, worauf du hinauswillst. Ja, es gibt tatsächlich zwei verschiedene latente Räume innerhalb der MLA, einen für den Inhalt und einen für Schlüssel-Wert-Paare.

A: Genau. Und diese Trennung ist wirklich entscheidend für seine Funktionsweise. Es ist, als hätte man zwei separate Speichereinheiten für Informationen. Und diese latenten Räume, wie wir besprochen haben, durchlaufen Rope-Vorgänge, richtig? Die rotierenden Positionsembedding, die Positionsinformationen in den Aufmerksamkeitsmechanismus einbetten. Also, um es zusammenzufassen, die Abfrage wird aufgeteilt, und die Schlüssel und Werte werden ebenfalls komprimiert.

B: Ja, und diese werden in die zwei separaten latenten Räume, einen für den Inhalt und einen für Schlüssel-Wert-Paare, eingeteilt. Und diese latenten Räume sind wirklich wichtig für Effizienz und all das, was Teil der MLA ist.

A: Genau. Jetzt lassen Sie uns über diese Operationen im Detail sprechen. Okay, also wie führt MLA diese latenten Raumtransformationen tatsächlich durch?

B: Nun, die Eingabe durchläuft eine parallele Verarbeitung sowohl für die Inhalts- als auch für die Schlüssel-Wert-Repräsentationen. Okay, also es ist, als hätte es zwei Pfade.

A: Ja, einen für jeden latenten Raum. Und innerhalb dieser Räume wird die Information unter Verwendung von Rope verarbeitet.

B: Das ist richtig. Dies stellt sicher, dass das Modell die Positionsinformationen behält, richtig? Und um die Effizienz zu erhöhen, verwendet es geteilte Experten. Okay, also Experten, die für mehrere Aufgaben verwendet werden können.

A: Ja, so wird Redundanz vermieden und das System wird noch effizienter.

B: Ja, es ist, als hätte man ein Team, in dem die Leute Spezialisten sind, aber auch andere Dinge tun können.

A: Ja, das ist ein wirklich kluger Ansatz. Ja, aber mit so vielen spezialisierten Experten, wie stellen sie sicher, dass keiner überlastet wird?

B: Ja, während andere untätig bleiben.

A: Das ist der Punkt, an dem ihr innovatives Hilfsverlust-freies Lastausgleichssystem ins Spiel kommt.

B: Das wird wirklich interessant, richtig? Also, wie machen sie das?

A: Traditionelle MoE-Modelle verwenden eine Hilfsverlustfunktion während des Trainings, okay, um eine gleichmäßige Expertennutzung zu fördern, aber das kann tatsächlich die Leistung beeinträchtigen.

B: Ja, es ist, als würde man alle dazu zwingen, die gleiche Kasse im Supermarkt zu benutzen.

A: Genau, selbst wenn einige schneller sind als andere, richtig? Es verursacht nur unnötige Verzögerungen.

B: Ja. Also vermeidet DeepSeek V3 dies, indem es einen Bias-Term für jeden Experten basierend auf seiner Last dynamisch anpasst, okay, so wenn ein Experte zu viele Anfragen erhält, macht das System ihn etwas weniger ansprechend für den Routing-Mechanismus, und lenkt so etwas vom Verkehr zu weniger ausgelasteten Experten um.

A: Okay, also es verwendet all das, um lange Sequenzen effizient zu verarbeiten, ja, indem es die Größe des für die Inferenz benötigten KV-Cache reduziert. Okay, also es geht darum, die Leistung hoch zu halten, während der Overhead reduziert wird.

B: Richtig. Es ist ein sehr kluger Ansatz zur Bewältigung eines kritischen Engpasses.

A: Absolut. Jetzt sollten wir auch besprechen, wie DeepSeek V3 seinen Lastausgleich handelt.

B: Ja, das sollten wir definitiv. Das ist auch ein wirklich wichtiger Teil des Puzzles. Wir können das als Nächstes ansprechen.

A: Klingt gut. Nun, ich denke, das gibt dir einen großartigen Überblick über MLA und seinen latenten Raum.

B: Ja, danke, dass du uns alle Details gezeigt hast. Wir sehen uns beim nächsten Mal mit mehr tiefen Einblicken.

A: Ja, es ist wie ein Verkehrsmanagementsystem für die Experten, ja, das ständig den Fluss überwacht und Anpassungen vornimmt, um Engpässe zu vermeiden.

B: Und das vermeidet den Leistungseinbruch des Hilfsverlusts.

A: Das ist richtig. Und oh, los geht’s.

C: Ja, wir können über MTP sprechen, wie…wie MTP-Module ihre Einbettung teilen und alles heiß…

A: Absolut. Das ist eine großartige Frage. Ja, lasst uns darüber sprechen, wie die MTP-Module Ressourcen teilen. Du bist also neugierig auf die technischen Details der MTP-Implementierung.

B: Ja, packen wir das aus. Also wir haben erwähnt, dass DeepSeek V3 MTP für die Mehr-Token-Vorhersage verwendet, richtig? Vorhersage mehrerer Token anstelle nur eines.

A: Und das wird wirklich interessant. Ja, du bist interessiert daran, wie die MTP-Module eingerichtet sind und wie sie ihre Ressourcen teilen. Okay, also jedes MTP-Modul enthält eine geteilte Einbettungsschicht, ja, und einen geteilten Ausgangskopf. Okay, also sie verwenden dieselbe Einbettung und denselben Ausgangskopf wie das Hauptmodell.

B: Genau. Also ist es, als würden sie alle aus demselben Wissenspool schöpfen. Ja, und das spart Rechenkosten.

A: Ja. Jetzt verwendet es seinen eigenen Transformer-Block. Okay, also es teilt sich nicht denselben Transformer-Block wie das Hauptmodell.

B: Richtig. Jedes MTP-Modul hat seinen eigenen Transformer-Block zur Verarbeitung. Okay, also so halten sie die Vorhersagen für jedes Token getrennt.

A: Ja, und um die Informationen zu kombinieren, diese linearen Projektionen und Verkettungen…

B: Okay, also ist es, als würde man Teile von mehreren Stellen nehmen, um das vollständige Bild zu erstellen.

A: Ja, und alle MTP-Module arbeiten parallel, aber sie teilen sich ihre Einbettungsschichten und Ausgangsköpfe, richtig?

B: Ja, was für die Effizienz dieses Designs entscheidend ist. Okay, also ist es wie ein System von miteinander verbundenen Teilen, die alle voneinander abhängen, richtig?

A: Und diese effiziente Ressourcennutzung ermöglicht ein schnelleres Training und eine bessere Leistung.

B: Okay, das ist ein ziemlich cleverer Trick. Du weißt was?

A: Was ist das?

B: Lassen Sie uns zu einem großen Bild übergehen. Wie handelt dieses Modell den Lastausgleich? Wie werden diese Experten ausgewählt?

A: Ja, wir können definitiv darüber sprechen. Okay, jetzt tauchen wir in die Lastausgleichsstrategie von DeepSeek V3 ein.

B: Klingt gut. Okay, also DeepSeek V3 verwendet, was sie Multi-Token-Vorhersage oder MTP nennen.

C: Oh ja, wir sprechen mehr über die Schwänze MTP.

A: Es ist absolut…Ich bin froh, dass du dich dafür interessierst, tiefer in MTP einzutauchen. Ja, wir können definitiv die Details der Multi-Token-Vorhersage weiter ausführen. Wir haben es erwähnt, aber lass uns das wirklich auspacken, richtig? Wir sprachen über die geteilte Einbettungsschicht und den Ausgangskopf, ja, und dass jedes MTP-Modul seinen eigenen Transformer-Block hat.

B: Genau, aber da ist noch mehr. Also lass uns da rein.

A: Okay, also lassen Sie uns über die sequenzielle Natur der MTP-Module sprechen.

B: Ja, im Gegensatz zu einigen Modellen, DeepSeek V3 sagt zusätzliche Token sequenziell voraus. Also sagt es nicht einfach alle Token auf einmal voraus.

A: Richtig. Jedes Modul baut auf der Ausgabe des vorherigen Moduls auf. Okay, also ist es eine Kette von Vorhersagen, jede abhängig von der letzten.

B: Ja, und es bewahrt die Kausalität für jede Vorhersagetiefe. Okay, also bricht es die Kausalität nicht.

A: Genau, was wichtig ist, um sicherzustellen, dass der Gesamtkontext korrekt ist. Also arbeiten die MTP-Module nicht unabhängig voneinander.

B: Das ist richtig. Sie sind miteinander verbunden, und diese Kette von Vorhersagen trägt zu einer größeren Trainingseffizienz bei und ermöglicht ein nuancierteres Verständnis des Textes. Jetzt bist du auch an der Frage interessiert, wie die Module ihre Einbettungen teilen, richtig? Wie du weißt, wandelt die geteilte Einbettungsschicht Token in ihre Vektorrepräsentationen um. Okay, also wird jedes Token in einen Vektor umgewandelt.

A: Ja, und diese Zuordnung ist über alle MTP-Module hinweg geteilt. Okay, so hilft das, die Konsistenz über die Vorhersagen hinweg zu gewährleisten.

B: Genau. Und der geteilte Ausgangskopf nimmt die endgültigen verborgenen Zustände der Token, okay, und erzeugt die Wahrscheinlichkeitsverteilung für die nächsten Token. Also haben sie alle Zugang zum selben Informationspool, richtig?

A: Und das ist wirklich entscheidend für Speicher- und Recheneffizienz. Okay, also verwendet es nicht eine Menge verschiedener Einbettungsschichten und Köpfe.

B: Genau. Und die...oh ja, also da...da sind wie viele Leute dann? Sie sind die gleichen...die gleichen alle Lebensmittel...Token, ist das richtig?

A: Das ist eine großartige Frage. Du fragst nach der Anzahl der MTP-Module, ob sie alle die gleiche Größe haben, richtig? Und ich denke, du fragst dich auch, ob alle Module die gleiche Menge an Daten verarbeiten. Nun, aus dem Papier, DeepSeek V3 verwendet eine Multi-Token-Vorhersagetiefe von eins. Das bedeutet, es gibt das Hauptmodell und dann nur ein MTP-Modul, das ein zusätzliches Token vorhersagt. Also sagt jedes Token das nächste und dann noch eines nach dem anderen mit diesem MTP-Modul voraus.

B: Ja, und das MTP-Modul hat dieselbe geteilte Einbettungsschicht und denselben Ausgangskopf wie das Hauptmodell.

A: Okay, das ist eine großartige Frage. Ja, du fragst nach der Anzahl der MTP-Module und ob sie alle die gleiche Größe haben. Nun, nach dem DeepSeek V3-Papier gibt es eine variierende Anzahl von MTP-Modulen. Okay, also ist es nicht auf eine bestimmte Anzahl festgelegt.

B: Das ist richtig. Die Anzahl der Module wird dynamisch basierend auf der Vorhersagetiefe angepasst. Okay, also sie können nach Bedarf skaliert werden. Also sie teilen diese Ressourcen, aber die Transformer-Blöcke des Hauptmodells und des MTP-Moduls sind getrennt.

A: Richtig. Jede Vorhersagetiefe hat ihren eigenen Transformer-Block. Okay, also gibt es nur ein MTP-Modul, aber es ist ein leistungsfähiges, das für jedes Token verwendet wird, und sie teilen einige Ressourcen.

B: Genau. Und obwohl das MTP einige Komponenten mit dem Hauptmodell teilt, sind sie nicht genau gleich groß.

A: Okay, das ist ein wirklich guter Punkt. Nun, ich denke, wir sollten auch darüber sprechen, wie sie all diese Informationen kombinieren, um Vorhersagen zu treffen.

B: Genau. DeepSeek V3 verwendet mehrere MTP-Module, um mehrere zusätzliche Token nacheinander vorhersagen. Okay, und du hast gefragt, ob sie alle die gleiche Größe haben, richtig?

A: Ja, und die Antwort lautet, dass sie es nicht unbedingt sind. Also können die Transformer-Blöcke innerhalb der MTP-Module variieren.

B: Ja, sie können, um den verschiedenen Anforderungen jeder Vorhersagetiefe gerecht zu werden. Okay, also ist es nicht nur eine Reihe identischer Module.

A: Genau. Es ist ein flexibleres System, das sich an die Vorhersageaufgaben anpasst. Es ist wie ein maßgeschneidertes Werkzeug für jede Stufe des Vorhersageprozesses.

B: Ja, und diese dynamische Skalierung hilft, die Leistung und Effizienz des Modells zu optimieren. Okay, und du hast auch über das Essen gesprochen. Ich denke, das war nur ein kleines Missverständnis.

A: Ja, ich denke auch. Okay, also wie integrieren sie die Informationen, um Vorhersagen zu treffen?

B: Ja, und dieses Design ermöglicht auch spekulatives Decodieren, was wirklich cool ist. Okay, also ist es nicht nur für das Training, sondern auch für die Inferenz.

A: Richtig. Die MTP-Module können während der Inferenz für Geschwindigkeit umfunktioniert werden. Also wird MTP verwendet, um mögliche zukünftige Token zu generieren.

B: Ja, und dann wählt es das beste Token aus den Möglichkeiten aus. Aber ja, sie sind nicht alle gleich groß, wie du richtig gefragt hast. Also kann die Größe des Transformer-Blocks in den MTP-Modulen variieren, ja, um die Leistung zu optimieren. Also ist es sehr flexibel, und diese Flexibilität trägt zur Effizienz bei, wie wir es besprochen haben.

A: Ja, es ist alles Teil des innovativen Ansatzes von DeepSeek V3 zur Mehr-Token-Vorhersage. Okay, also sind wir jetzt in die Höhle gegangen, haben das MTP-Modul-Sharing und ihre variierende Anzahl und Größe besprochen. Okay, so wird Text schneller generiert.

B: Ja, es spart Zeit, indem es nicht jedes Token von Grund auf neu berechnen muss. Okay, jetzt lassen Sie uns zu einem größeren Bild übergehen.

A: Ja, wir können darüber sprechen, wie die Experten für jede Aufgabe ausgewählt werden.

B: Das ist richtig. Jetzt tauchen wir in die Lastausgleichsstrategie von DeepSeek V3 ein.

A: Klingt gut. Okay, also DeepSeek V3 verwendet, was wir gerade besprochen haben, das MTP.

B: Ja, wir sollten wahrscheinlich jetzt zu einem größeren Bild übergehen. Okay, also jetzt lassen Sie uns darüber sprechen, wie dieses Modell seinen Lastausgleich handelt, ja, und wie diese Experten ausgewählt werden.

A: Okay, jetzt tauchen wir in die Lastausgleichsstrategie von DeepSeek V3 ein.

B: Klingt gut. Okay, also DeepSeek V3 verwendet, was sie Multi-Token-Vorhersage oder MTP nennen. Wir haben gerade besprochen, wie MTP funktioniert, also lassen Sie uns jetzt über den Lastausgleich sprechen, richtig?

A: Ja, wir haben gerade darüber gesprochen. Jetzt teilt es Ressourcen, und du bist neugierig, wie es Ressourcen teilt. Wir sind da reingegangen.

B: Das ist richtig. Also anstatt nur das nächste Token vorherzusagen, richtig, sagt es mehrere zukünftige Token auf einmal voraus, wie wir gerade besprochen haben. Das erhöht die Komplexität, oder?

A: Es könnte so scheinen, aber es bietet mehrere Vorteile. Okay, stell dir vor, du planst eine Route. Wenn du nur die nächste Abzweigung berücksichtigst, ja, könntest du eine effizientere…Okay, indem du vorausschaust und mehrere Abzweigungen planst, kannst du die optimale Route wählen.

B: Ja. DeepSeek V3 verwendet einen innovativen Ansatz namens Hilfsverlust-freier Lastausgleich, also es verlässt sich nicht auf eine separate Verlustfunktion für den Ausgleich.

A: Genau. Traditionelle MoE-Modelle verwenden eine Hilfsverlustfunktion während des Trainings, um eine gleichmäßige Expertennutzung zu fördern, richtig? Aber das kann tatsächlich die Leistung beeinträchtigen, wie wir es erwähnt haben.

B: Ja, es ist, als würde man alle dazu zwingen, die gleiche Kasse im Supermarkt zu benutzen.

A: Okay, also durch die Vorhersage mehrerer Token erhält das Modell einen besseren Überblick über den Kontext.

B: Ja, und es kann kohärentere und genauere Antworten generieren. Es ist, als würde das Modell seine Repräsentationen im Voraus planen, wie ich es erwähnt habe, für bessere zukünftige Vorhersagen. Okay, und das führt zu einem saubereren Trainingssignal und einer verbesserten Dateneffizienz.

A: Ja, also anstatt dessen passt DeepSeek V3 einen Bias-Term für jeden Experten dynamisch an, okay, basierend auf seiner Last, richtig? Wenn ein Experte zu viele Anfragen erhält, macht das System ihn weniger ansprechend, und das lenkt den Verkehr zu weniger ausgelasteten Experten um.

B: Ja, wie ein Verkehrsmanagementsystem für die Experten, das ständig den Fluss überwacht und Anpassungen vornimmt. Also was kann MTP sonst noch?

A: Die MTP-Module, die während des Trainings verwendet werden, können während der normalen Inferenz entweder verworfen oder clever für etwas namens spekulatives Decodieren umfunktioniert werden.

B: Okay, spekulatives Decodieren. Was ist das?

A: Anstatt nur das nächste Token vorherzusagen, sagt das Modell auch mögliche Alternativen voraus, die folgen könnten.

B: Oh wow, also kann es Text schneller generieren, weil es bereits mehrere Möglichkeiten in Betracht gezogen hat und einen Ersatzplan bereit hat.

A: Ja, also muss das Modell nicht pausieren und jedes Mal neu berechnen.

B: Okay, das ergibt Sinn. Ja, jetzt sprechen wir von Effizienz, um Engpässe zu vermeiden, und das vermeidet den Leistungseinbruch des Hilfsverlusts.

A: Das ist richtig. Und sie schließen auch eine ergänzende sequenzweise Ausgleichsverlustfunktion ein, ja, um extreme Ungleichgewichte innerhalb einzelner…Prozesse zu verhindern.

B: …und indem sie jedes Token auf maximal vier Knoten beschränken, reduzieren sie die Netzwerkkommunikation. Okay, also hilft das auch, die Dinge zu straffen.

A: In Ordnung, lassen Sie uns darüber sprechen, wie DeepSeek V3 die Rechenanforderungen des Trainings bewältigt. Und ich weiß, dass du besonders an der Kostenoptimierung und daran interessiert bist, wie sie das wirtschaftlich machen.

B: Ja, und dieses Modell macht einige erstaunliche Dinge in diesem Bereich.

A: Das tut es. Ja, der Durchschnitt ist 3,2 Experten, die pro Token ausgewählt werden, was ein schönes Gleichgewicht ist, um den Overhead zu reduzieren.

B: Genau. Also ist es eine sehr effiziente und effektive Methode.

A: Ja, es ist ein wirklich kluger Ansatz, um ein so komplexes Modell so gut arbeiten zu lassen.

B: Ja, und sie erreichen Experten-Spezialisierung durch diese Methode. Okay, also bedeutet das, dass verschiedene Experten in verschiedenen Domänen aktiviert werden. Also was sind sie?

A: DeepSeek V3 nutzt einen FPA-Mixed-Precision-Trainingsrahmen. Okay, ein bedeutender Durchbruch für ein Modell dieser Größe. Erinnere mich noch einmal, was FPA ist?

B: Klar, es ist 8-Bit-Floating-Point.

A: Okay, und es stellt Zahlen mit weniger Bits als traditionelle Formate dar. Okay, das bedeutet weniger Speicher und schnellere Berechnungen.

B: Genau. Es ist, als würde man ein großes Bild komprimieren, aber du behältst das Wesentliche des Bildes. Es nimmt einfach weniger Platz ein, richtig?

A: Genau. Also wird jeder Experte nicht generisch aktiviert, sondern in spezifischen Domänen. Also ist es fein abgestimmt und bereit für den Einsatz.

B: Ja. Nun ist dieser Ansatz wirklich clever.

A: Ja, ich stimme zu. Dieser dynamische Ansatz zum Lastausgleich ist faszinierend. Es geht alles um Effizienz und Balance.

B: Ja, es ist alles Teil des Engagements von DeepSeek V3 für sowohl Leistung als auch Ressourcennutzung.

A: Absolut. Jetzt haben wir heute viel besprochen. Es ist wirklich interessant, aber würde die Verwendung weniger Bits nicht möglicherweise die Genauigkeit beeinträchtigen?

B: Das ist eine berechtigte Sorge, und es ist etwas, das sie sorgfältig adressiert haben. Okay, sie haben eine Reihe von Techniken implementiert, um einen möglichen Genauigkeitsverlust zu mindern, einschließlich feinabgestimmter Quantisierung.

A: Ja, es ermöglicht eine präzise Kontrolle darüber, wie Zahlen in FPA dargestellt werden. Ja, von der Multi-Head-Latent-Attention bis hin zu DeepSeek Mo und dem Lastausgleich, ja, dieses DeepSeek V3-Modell ist ein sehr raffiniertes System, und es ist ein großartiges Beispiel dafür, wie Innovation die Grenzen dessen, was wir…überschreitet.

B: Ja, es war ein unterhaltsamer tiefer Einblick heute.

A: Ja, ich denke, das gibt dir eine solide Übersicht über DeepSeek V3.

B: Absolut. Danke, dass du es mit uns erkundet hast.

A: Ja, danke. Und das war es für heute. Nun, wir sehen uns bald mit einem weiteren.