XAI@DNUG

2025-11-18

CompanyGPT:

🍪 **Bob** sollte unbedingt die **Kapazitäten zur Produktion von Plätzchen** 🏭 erhöhen und eine **zugehörige Marketingkampagne** 📢 starten, da die **Nachfrage steigen wird** 📈 und der **Umsatz im Anschluss durch die Decke geht** 🚀 💰 .

Me@Work:

??? Warum sollte das passieren?

CompanyGPT:

Because I said so. 🤖 💬

XAI@DNUG

Erklärbarkeit in der KI, oder die Antwort auf die Frage

Warum!?

2025-11-18
Sven Flake
sflake@paiqo.com

Das alles ist „Künstliche Intelligenz“

(im heutigen Beispiel)

Vorhersagen (ML)

Empfehlungen (LP)

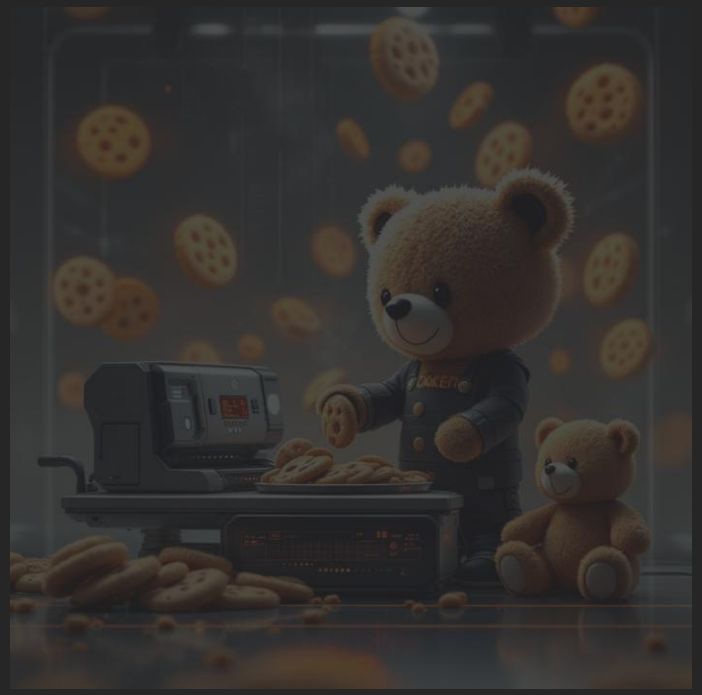
Generative KI (LLM)

Das alles ist „Künstliche Intelligenz“

(im heutigen Beispiel)

Vorhersagen (ML)

Empfehlungen (LP)



Generative KI (LLM)

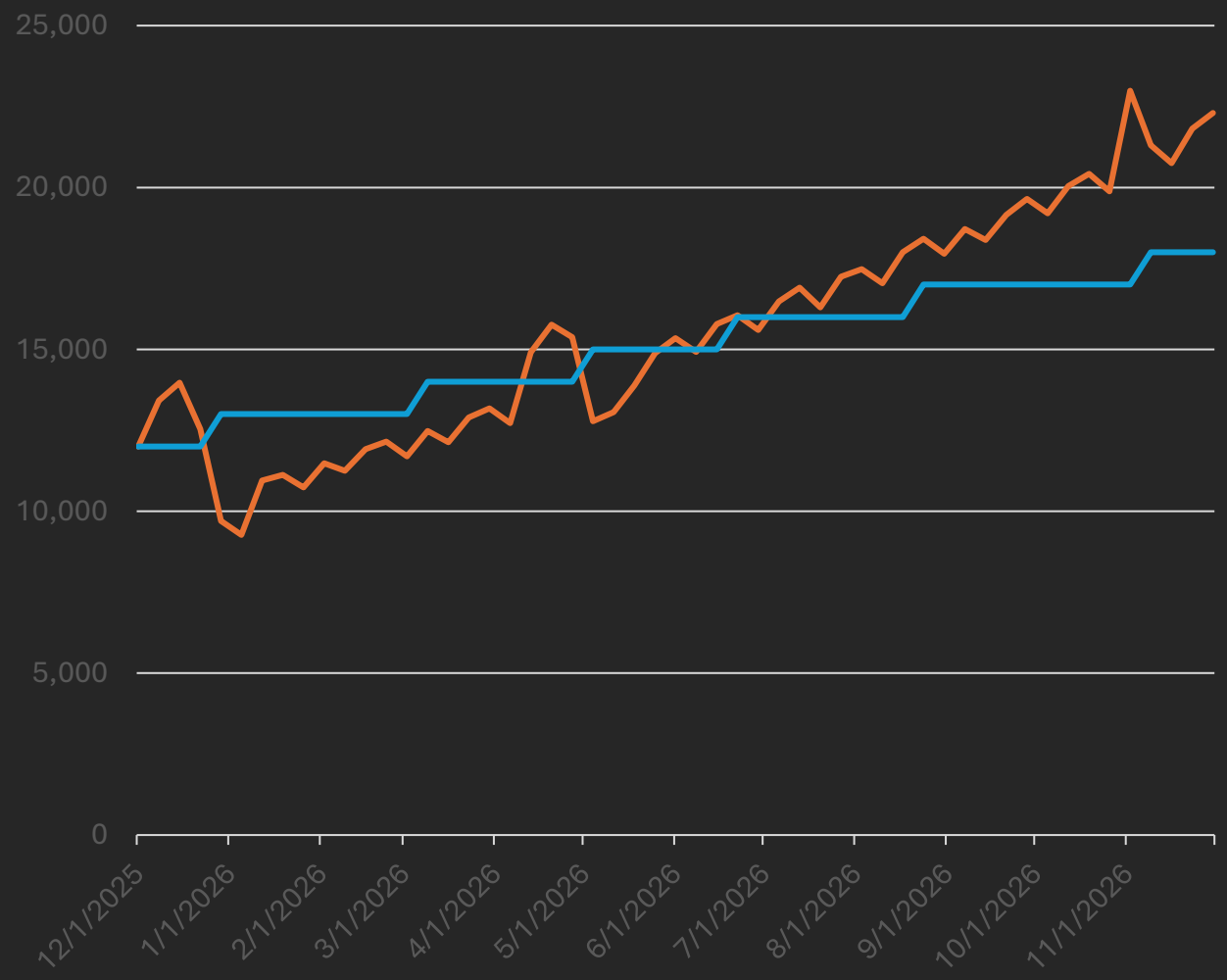
Backzeit!

- Bob hat ein Forecast-Modell für sein Geschäft erstellen lassen
- Das Modell sagt Umsätze im Plätzchengeschäft voraus
- Dagegen legt er die maximale Produktionskapazität

→ Der Barf liegt schon bald über seinen Kapazitäten für die Produktion

→ Kann er diesem Forecast trauen, um Investitionsentscheidungen zutreffen?

Warum sieht das Modell einen wachsenden Bedarf an Plätzchen?



Grundlagen

- Forecasts trainieren die Vorhersauge auf historischen Daten
- Forecasts stellen Zusammenhänge von Eingangsgrößen (Features) zu Zielgrößen her

→ Welche Features haben welchen Einfluss auf die Vorhersage?

Datum	Preis	Wetter	Marketing	Events	BIP	RKL	Absatz
2025-01-01	0.062736	0.499923	0.001182	1	0.101145	0.544345	962.6797
2025-01-02	0.06449	0.523635	0	0	0.082436	0.536846	943.3536
2025-01-03	0.088574	0.564595	0.00601	0.5	0.128449	0.599701	945.8849
2025-01-04	0.137528	0.575713	0.012517	1	0.081302	0.500088	963.4035
2025-01-05	0.136557	0.456917	0.013111	1	0.110334	0.582312	955.1706
2025-01-06	0.135588	0.493308	0.026468	1	0.073511	0.605496	952.9186
2025-01-07	0.186138	0.584405	0.019716	1	0.117427	0.593331	951.6534
2025-01-08	0.213624	0.596811	0.02397	0.5	0.057629	0.597138	939.9041
2025-01-09	0.205968	0.609266	0.031349	1	0.010471	0.611935	948.8131
2025-01-10	0.227065	0.802004	0.043417	1	0	0.64776	944.272
2025-01-11	0.219581	0.636747	0.05783	0	0.035572	0.575976	931.7424
2025-01-12	0.212031	0.679327	0.050399	1	0.040704	0.706908	939.0479

Diskurs: Spieltheorie

- Bob, Hermann und Oskar wollen die Marktführerschaft weihnachtlicher Backwaren
- Jeder hat Stärken und Schwächen
- Sie tun sich zusammen, um gemeinsam eine Erfolgsstrategie zu entwickeln

→ Wie teilen sie am Ende den Gewinn?

Diskurs: Spieltheorie

- Prüfe den Gewinn der Gruppe **ohne** eine Person
- Vergleiche mit dem Gewinn der Gruppe **mit** der Person

→ Persönlicher Beitrag zu dieser Gruppe

- Berechne dies für alle Personen und alle möglichen Koalitionen
- Bilde den Durchschnitt

→ Shapley-Wert der Person (nach Lloyd Shapley)

→ Der eigene Anteil am Gesamtbeitrag legt den jeweiligen Gewinnanteil fest

Zurück zum Forecast

- Gewinn → Vorhersage
- Teilnehmer → einzelne Features
- Gruppe → Menge der Features

Datum	Preis	Wetter	Marketing	Events	BIP	RKL	Absatz
2025-01-01	0.062736	0.499923	0.001182	1	0.101145	0.544345	962.6797
2025-01-02	0.06449	0.523635	0	0	0.082436	0.536846	943.3536
2025-01-03	0.088574	0.564595	0.00601	0.5	0.128449	0.599701	945.8849
2025-01-04	0.137528	0.575713	0.012517	1	0.081302	0.500088	963.4035
2025-01-05	0.136557	0.456917	0.013111	1	0.110334	0.582312	955.1706
2025-01-06	0.135588	0.493308	0.026468	1	0.073511	0.605496	952.9186
2025-01-07	0.186138	0.584405	0.019716	1	0.117427	0.593331	951.6534
2025-01-08	0.213624	0.596811	0.02397	0.5	0.057629	0.597138	939.9041
2025-01-09	0.205968	0.609266	0.031349	1	0.010471	0.611935	948.8131
2025-01-10	0.227065	0.802004	0.043417	1	0	0.64776	944.272
2025-01-11	0.219581	0.636747	0.05783	0	0.035572	0.575976	931.7424
2025-01-12	0.212031	0.679327	0.050399	1	0.040704	0.706908	939.0479

Interpretation

- Welchen Einfluss haben die Features auf den Forecast an einem speziellen Datum?

→ Einzelne Features erhöhen den Forecast

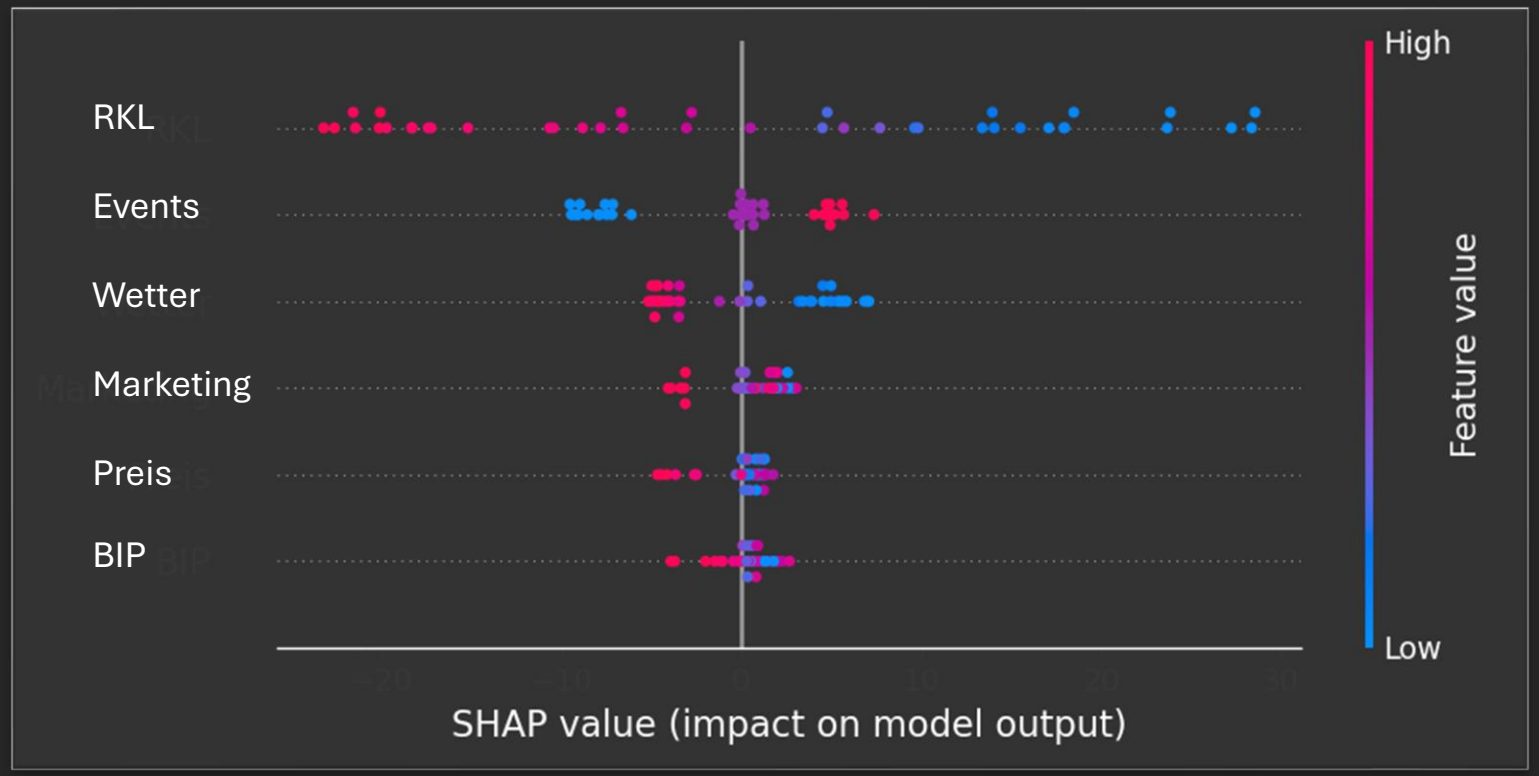
→ Einzelne Features verringern den Forecast

Interpretation

- Welchen Einfluss haben die Features insgesamt auf den Forecast?

→ Einzelne Features haben starke und wechselnde Einflüsse

→ Andere Features haben wenig oder einseitigen Einfluss



Wrap-up Forecasting

- Einfluss einzelner Features auf den Forecast lässt sich berechnen
- Möglich: Plausibilisierung
- Möglich: Bewertung von Maßnahmen

→ Insgesamt lässt sich ein Forecast so besser verstehen

Bob kann jetzt besser beurteilen, ob er der Vorhersage der Absatzsteigerung vertrauen kann

Das alles ist „Künstliche Intelligenz“

(im heutigen Beispiel)

Vorhersagen (ML)

Empfehlungen (LP)

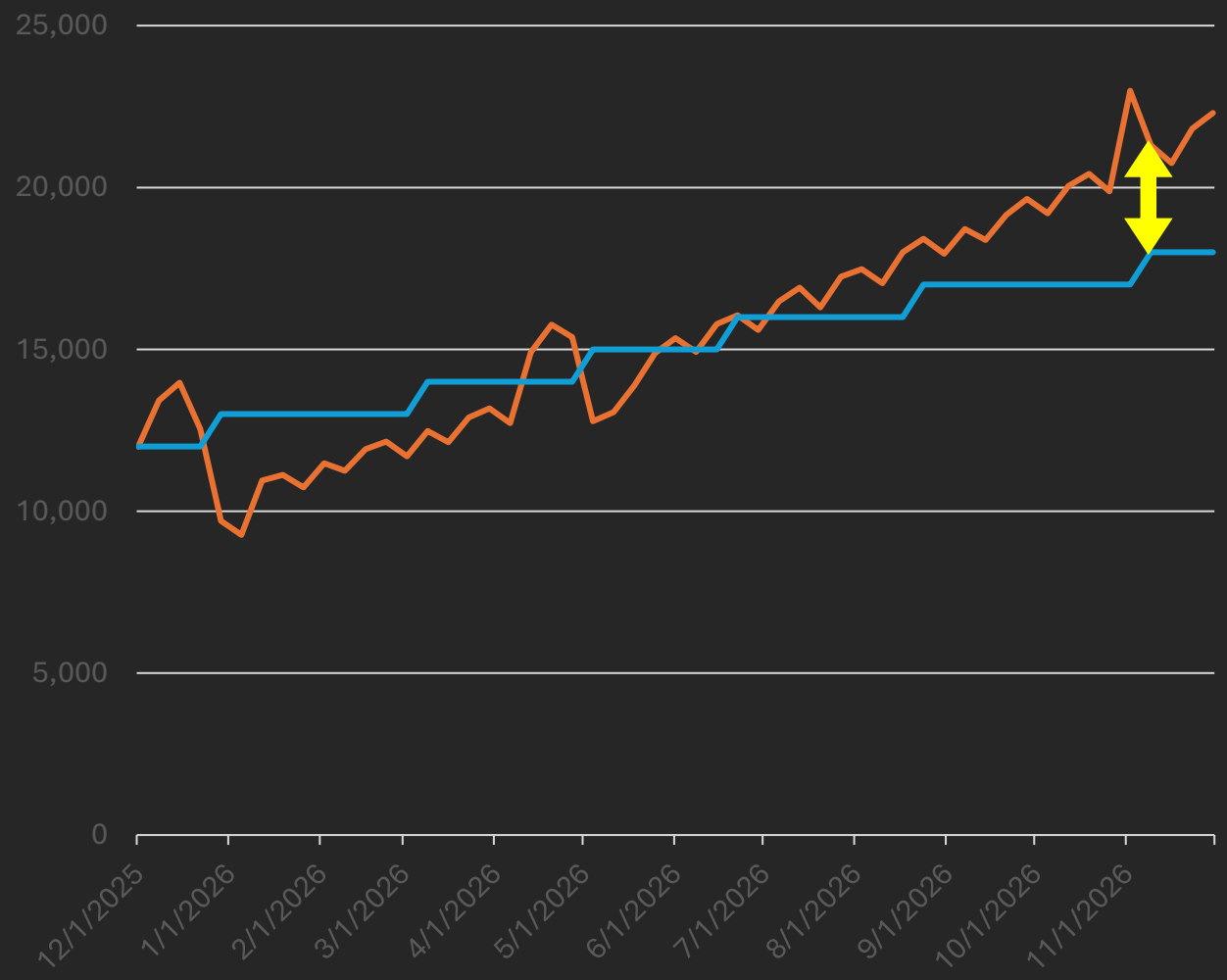
Generative KI (LLM)

Backzeit!

- Das Modell sagt Umsätze im Plätzchengeschäft voraus
- Dagegen legt er die maximale Produktionskapazität
- Die Produktion ist bereits optimal ausgelastet

→ Die Ressourcen reichen nicht, um den wachsenden Absatz zu befriedigen

Was sollte Bob tun, um die Produktion möglichst effektiv zu erweitern?



Plätzchen oder Kekse?

- Bob stellt insbesondere Plätzchen und Kekse her
- Mit beiden macht er unterschiedlichen Umsatz
- Beides benötigt unterschiedliche Zutaten und unterschiedlich viel Zeit
- Dabei werden begrenzte Ressourcen berücksichtigt
- Ein Optimierungsmodell schlägt vor, welche Mengen von welcher Sorte Bob herstellen sollte, um den Umsatz zu maximieren

$$k, p \geq 0$$

$$0{,}4k + 0{,}3p \leq 100$$

$$0{,}3k + 0{,}5p \leq 120$$

$$0{,}25k + 0{,}2p \leq 70$$

$$\max 7k + 8{,}5p$$

127 **163**

2025-11-18, Sven Flake, sflake@paiqo.com

Mehr von allem?

- Ziel: Mehr produzieren
- Ansatz: von allen Ressourcen mehr kaufen
- Aber: Gewürze sind teuer

→ Prüfe, was es im Umsatz bringt, die Ressourcen zu erhöhen

→ **Grenzkosten** der Ressourcen berechnen (**Duales Problem** mit **Schattenpreisen**)

$$k, p \geq 0$$

$$0{,}4k + 0{,}3p \leq 100$$

$$0{,}3k + 0{,}5p \leq 120$$

$$0{,}25k + 0{,}2p \leq 70$$

$$\max 7k + 8{,}5p$$

2025-11-18, Sven Flake, sflake@paiqo.com

Mehr von allem?

- Ziel: Mehr produzieren
- Ansatz: von allen Ressourcen mehr kaufen
- Aber: Gewürze sind teuer

→ Prüfe, was es im Umsatz bringt, die Ressourcen zu erhöhen

→ **Grenzkosten** der Ressourcen berechnen (**Duales Problem** mit **Schattenpreisen**)

2025-11-18, Sven Flake, sflake@paiqo.com

Mehr von allem?

- Ziel: Mehr produzieren
- Ansatz: von allen Ressourcen mehr kaufen
- Aber: Gewürze sind teuer

→ Prüfe, was es im Umsatz bringt, die Ressourcen zu erhöhen

→ **Grenzkosten** der Ressourcen berechnen (**Duales Problem** mit **Schattenpreisen**)

Umsatzerhöhung pro +1 kg

Vorrat an Gewürz wird nicht ausgereizt

8,64 € **11,82 €** **0,00 €**

$$0{,}4m + 0{,}3z + 0{,}25g \geq 7$$

$$0{,}3m + 0{,}5z + 0{,}2g \geq 8{,}5$$

$$m, z, g \geq 0$$

$$\min 100m + 120z + 70g$$

Direkte Beurteilung

$$k, p \geq 0$$

$$0{,}4k + 0{,}3p \leq 100$$

$$0{,}3k + 0{,}5p \leq 120$$

$$0{,}25k + 0{,}2p \leq 70$$

$$\max 7k + 8{,}5p$$

2025-11-18, Sven Flake, sflake@paiqo.com

Direkte Beurteilung

- Investitionsentscheidungen kann man direkt in das Modell einbauen
- Über „Soft Constraints“ werden Erweiterungen der Ressourcen bepreist

→ Das Modell entscheidet direkt bei der Lösung, ob es sich lohnt, Ressourcen zuzukaufen

$$k, p, m, z, g \geq 0$$

~10

$$0{,}4k + 0{,}3p \leq 100 + m$$

$$0{,}3k + 0{,}5p \leq 120 + z$$

$$0{,}25k + 0{,}2p \leq 70 + g$$

$$\max 7k + 8{,}5p - f_m m - f_z z - f_g g$$

169

138

2025-11-18, Sven Flake, sflake@paiqo.com

Wrap-up Optimierung

- Mit harten Constraints kann man nicht nur Ressourcen beschränken, sondern auch ihren Wert beurteilen
- Mit weichen Constraints kann man sinnvolle Investitionsentscheidungen direkt ins Modell bauen

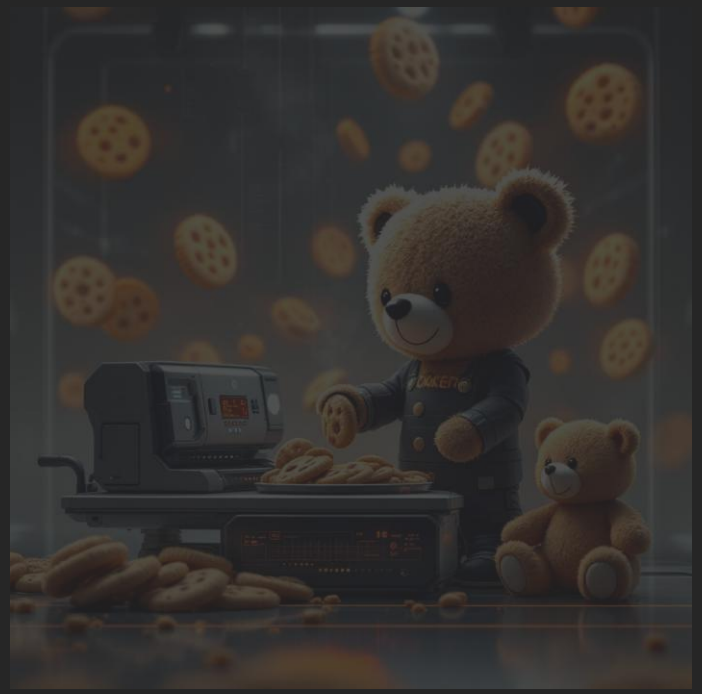
Bob kann sich von seiner Produktionsplanung direkt vorschlagen lassen, ob sich Investitionen rechnen

Das alles ist „Künstliche Intelligenz“

(im heutigen Beispiel)

Vorhersagen (ML)

Empfehlungen (LP)



Generative KI (LLM)

Der Bro weiß alles

- Bobs Chatbot macht sehr konkrete Handlungsvorschläge
- Genauso kann er zu zentralen Unterlagen berichten, Informationen zusammenstellen oder als Coach fungieren

→ Aber warum sollten wir dem Ding vertrauen?

Grundproblem

- Auch LLMs müssen bieten:
  - Vertrauen
  - Auditierbarkeit
  - Fehlersuche
  - Sicherheit

Aber:

- LLMs generieren keine Ergebnisse nach festen Regeln
- Es gibt keine klaren Features
- Fähigkeiten sind emergent
- Parameter sind im Milliardenbereich

Ansatz 1: Mechanistisch

Ziel:

- Welcher Bestandteil kann was?

Idee:

- Modell in Bestandteile zerlegen: Neuronen, Layer, Attention-Heads, ...

Beispiel:

- Bestimmte Teile eines Layers erkennen semantische Nähe

Vorteile:

- Präzise, gute Einsichten in Zusammenhänge, Grundlage für Safety-Analysen

Nachteile:

- Extrem aufwändig, skaliert schlecht, Überinterpretation

Ansatz 2: Attribution

Ziel:

- Bedeutung der Eingaben verstehen

Idee:

- Welche Eingaben haben den größten Einfluss auf die Ausgabe?

Beispiel:

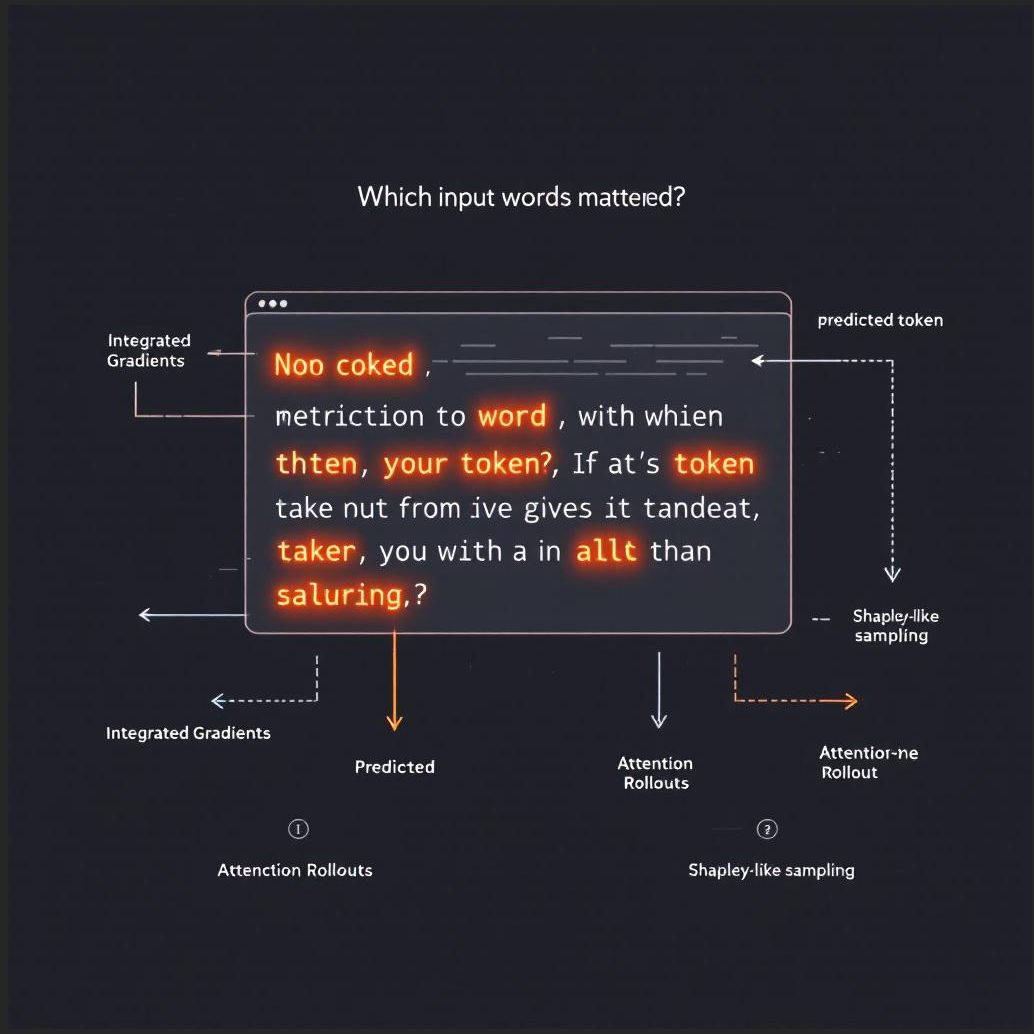
- Verdeutlichung der Eingabegewichtung macht Ergebnis nachvollziehbar

Vorteile:

- Gut visualisierbar, unterstützt Prompt-Engineering, verständlich

Nachteile:

- Nicht kausal, Tokenisierung beeinflusst Interpretation



Ansatz 3:
Rational Generation

Ziel:

- Das Modell zu einer logischen Kausalkette zwingen

Idee:

- Das Modell erklärt sich selbst

Beispiel:

- Schrittweise Erläuterung, warum Paris die Hauptstand von Frankreich ist

Vorteile:

- Verbessert Ergebnis, sehr nutzerfreundlich

Nachteile:

- Nicht kausal, nicht automatisch wahr, bildet nicht das tatsächliche Reasoning ab

Ansatz 4: Behavioral Testing

Ziel:

- Verhaltenskonsistenz prüfen

Idee:

- Das LLM im Verhaltenslabor

Beispiel:

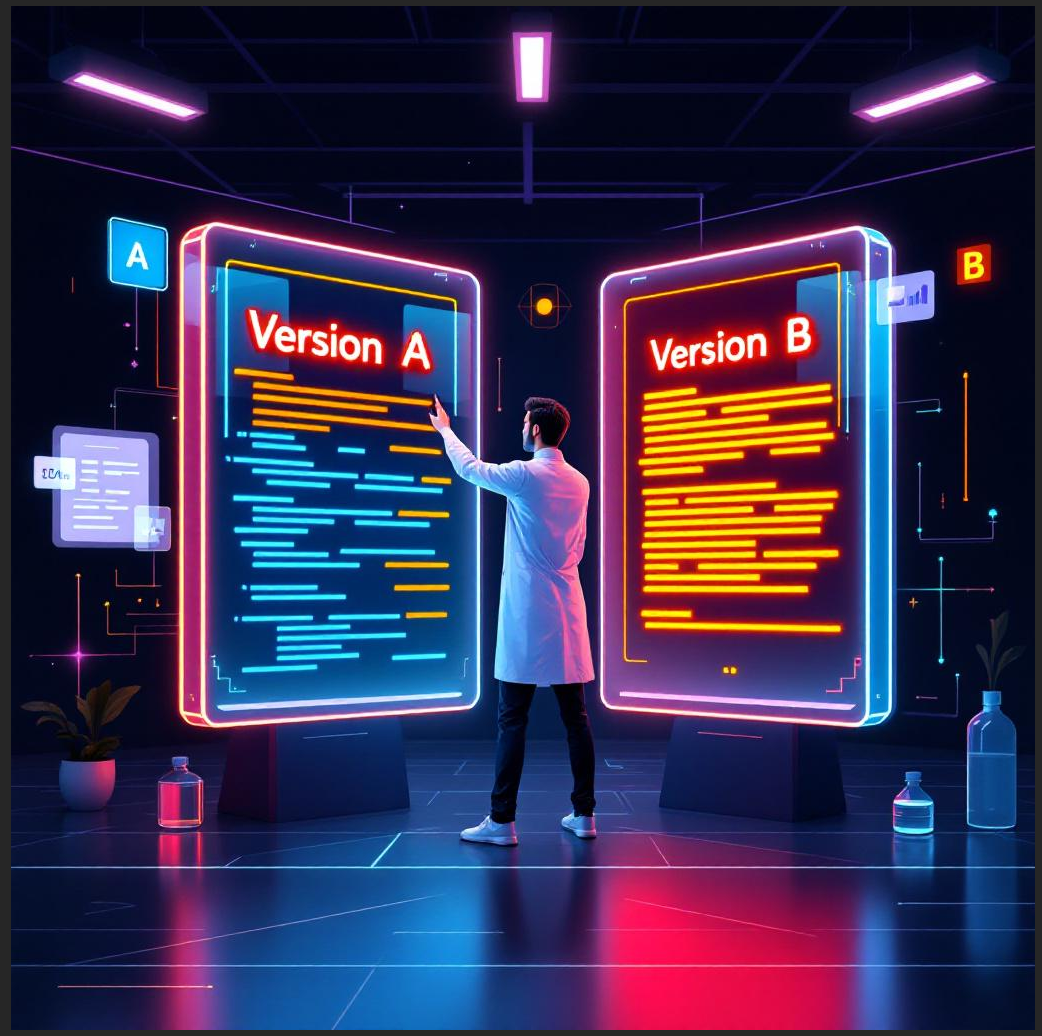
- „Warum braucht der Entwickler mehr Zeit?“ vs. „Warum braucht die Entwicklerin mehr Zeit?“

Vorteile:

- Skaliert gut, aussagekräftig für Sicherheit und Fairness

Nachteile:

- Liefert keine Einsichten in interne Mechanismen, liefert keine Erklärungen



Wrap-up Generative KI

- Derzeit kein Ansatz, der vergleichbar mächtig ist wie Shapley/Shadow Pricing bei bewährten Modellen
- Vielversprechend sind Kombinationen, heute im Beispiel:
  - Mechanistische Interpretation
  - Attribution
  - Rationale Generation
  - Behavioral Testing

→ Trotzdem unklar: Werden wir echte Erklärbarkeit erreichen?

Wrap-up

Vorhersagen (ML) ✓ **Empfehlungen (LP)** ✓ **Generative KI (LLM)** ?

Was bedeutet das für unseren zukünftigen Umgang mit Agenten?