Analyse der KI-Plattform Performance bei Pizza Napoli Rezepten

Erstellt mit ChatGPT o3 24.7.2025 durch Ali Soy

Inhaltsverzeichnis

1. Mengenvergleich der Zutaten *3*

Teig-Zutaten 3

Saucen-Zutaten 5

Belag-Zutaten 8

2. Vollständigkeit und Übereinstimmung der Zutaten *12*

3. Qualitative Auffälligkeiten *15*

4. Bewertung der KI-Plattformen *19*

Zusammenfassung *22*

Mengenvergleich 22

Vollständigkeit 22

Qualitative Auffälligkeiten 23

Bewertung 23

Fazit 23

Sources *24*

Prompt für den Report *24*

1. Mengenvergleich der Zutaten

Zunächst wurden alle **Zutatenmengen auf einheitliche Einheiten** gebracht, basierend auf den Angaben in der Spalte *Zusatzinfo*. Beispielsweise wurden Teelöffel (TL) und Esslöffel (EL) in Gramm bzw. Milliliter umgerechnet (z.B. 1 TL Salz = 5 g, 1 EL Olivenöl = 15 ml). Dadurch lassen sich die von den KI-Plattformen vorgeschlagenen Mengen direkt mit dem hochgerechneten Betty Bossi Referenzrezept (für 503 Personen) vergleichen.

Teig-Zutaten

In der folgenden Tabelle sind die Mengen der Hauptteig-Zutaten (für 503 Personen) aufgeführt – Referenz vs. KI-Plattformen. Alle Werte wurden in Kilogramm (kg), Litern (l) bzw. Gramm (g) vereinheitlicht:

Zutat	Referenz (BB)	ChatGPT 4	ChatGPT 3.5	Perplexity	DeepSeek	Grok	Claude
Mehl (kg)	62,9	63,0	78,1	88,0	62,9	79,5	12,5
Wasser (l)	40,7	38,0	50,7	53,0	37,7	47,7	8,3
Salz (Teig, kg)	0,96	1,60	1,95	1,50	1,26	2,39	0,25
Hefe (g)	1268	500	40	50	1884	2390	250
Olivenöl (Teig, l)	3,8	2,5	0	0	3,0	4,0	0,25

Tab.: Mengenvergleich der Teig-Zutaten (für 503 Portionen) – Betty Bossi Referenz vs. KI-Plattformen.

- **Mehl:** Die Referenzmenge (62,9 kg) liegt nahe am Durchschnitt der KI-Vorschläge (ca. 64,0 kg). Die Spannweite ist jedoch groß – von **nur 12,5 kg** (Claude) bis **88,0 kg** (Perplexity). Die meisten KI-Plattformen lieferten Mehlmengen im gleichen Grössenordnungsbereich wie die Referenz (zwischen ~63–79 kg), mit Ausnahme von Claude (nur 12,5 kg, ein extremer Ausreisser nach unten) und Perplexity (88 kg, deutlich höher). Die Standardabweichung beträgt ca. 24,7 kg, was eine beträchtliche **Streuung** der Mehlmengen anzeigt.

Gegenüber dem Referenzwert weichen die meisten KI-Mengen um ±0–25% ab – **Perplexity liegt ~40% darüber**, während Claude über 80% darunter liegt.

- **Wasser:** Im Schnitt gaben die KIs **39,2 l** Wasser an (Referenz: 40,7 l). Die Spannweite reicht von **8,3 l** (Claude) bis **53,0 l** (Perplexity). Die Standardabweichung (~15,0 l) ist hoch. Die meisten Werte (38–53 l) liegen in einem plausiblen Bereich um den Referenzwert, abgesehen von Claude, dessen Wassermenge (8,3 l) extrem niedrig ausfiel (nur ~20% der Referenz). Perplexity schlug mit 53 l etwas mehr Wasser vor (+30% gegenüber Referenz), ChatGPT 3.5 und Grok lagen moderat darüber (~25% mehr), während ChatGPT 4 und DeepSeek leicht weniger Wasser angaben (~5–7% unter Referenz). **Insgesamt bewegten sich die Wasserangaben (Claude ausgenommen) innerhalb ±30% der Referenzmenge.**
- **Salz (für den Teig):** Die Referenz enthält ~0,96 kg Salz im Teig. Hier zeigten die KI-Angaben deutliche Abweichungen: **Grok verwendete mit 2,39 kg am meisten Salz** (mehr als doppelt so viel wie Referenz), gefolgt von ChatGPT 3.5 (1,95 kg) und ChatGPT 4 (1,60 kg). Perplexity (1,50 kg) und DeepSeek (1,26 kg) lagen moderat über dem Referenzwert (+30% bzw. +15%). Claude fällt wieder stark aus dem Rahmen mit nur 0,25 kg Salz (rund ein Viertel der benötigten Menge). Im Durchschnitt wurden ~1,49 kg Salz vorgeschlagen, was etwa **55% über dem Referenzwert** liegt. Die Spannweite (0,25–2,39 kg) und die Standardabweichung (~0,66 kg) zeigen grosse Unterschiede. **Einige KIs salzen den Teig also erheblich stärker als die Referenz, andere – insbesondere Claude – viel zu wenig.**
- **Hefe:** Die Unterschiede bei der Hefe sind besonders ausgeprägt. Die Betty Bossi Referenz (linear skaliert) würde ca. **1268 g** Hefe erfordern. Die KI-Plattformen wichen hier teils extrem ab. **ChatGPT 3.5 und Perplexity gaben mit 40 g bzw. 50 g verschwindend geringe Hefemengen an**, was nur ~3–4% der Referenzmenge entspricht – praktisch unzureichend für die Teigmenge. ChatGPT 4 schlug 500 g Hefe vor (~40% der Referenz), Claude 250 g (~20%). Demgegenüber übertrafen **DeepSeek (1884 g)** und **Grok (2390 g)** den Referenzwert (≈1268 g) deutlich (+50% bzw. +88%). Der Mittelwert liegt bei ~852 g, doch die Standardabweichung (≈933 g) übersteigt den Mittelwert, ein Indikator für **sehr grosse Schwankungen**. **Keine KI traf die Referenzmenge – alle lagen entweder deutlich darunter oder darüber.** Manche KI-Rezepte (v.a. GPT 3.5 und Perplexity) enthalten so wenig Hefe, dass der Teig in der Praxis

kaum aufgehen würde, während Grok und DeepSeek sehr hohe Hefemengen ansetzen.

- **Olivenöl im Teig:** Das Referenzrezept enthält ~3,8 l Öl im Teig (entspricht 0,5 EL pro Portion). Die KI-Angaben variieren: **Zwei Plattformen (ChatGPT 3.5 und Perplexity) verzichten komplett auf Öl im Teig**, wohingegen Grok mit ~4,0 l sogar leicht mehr als die Referenz nutzt. DeepSeek (3,0 l) und ChatGPT 4 (2,5 l) verwenden etwas weniger Öl als die Referenz. Claude gibt mit 0,25 l wieder ein Extrem nach unten an (nur ~7% der Referenzmenge). Durchschnittlich wurden ~1,6 l Öl genannt, aber dieser Wert ist wenig aussagekräftig, da die Streuung gross ist (Std ~1,6 l, Spannweite 0–4,0 l). **Hier fällt auf, dass einige KIs – evtl. aus Authentizitätsgründen – gar kein Öl im Pizzateig verwenden (was bei original neapolitanischem Teig tatsächlich vorkommt), während andere ähnliche Grössenordnungen wie das Referenzrezept vorschlagen.**

Saucen-Zutaten

Auch bei den Zutaten für die Tomatensauce treten markante Unterschiede zutage. Unten eine Übersichtstabelle der wichtigsten Saucenkomponenten (alle Mengen für 503 Portionen, vereinheitlicht):

Zutat	Referenz (BB)	ChatGPT 4	ChatGPT 3.5	Perplexity	DeepSeek	Grok	Claude
Zwiebel (kg)	18,9	0	0	0	0	0	0
Olivenöl (Sauce, l)	1,9	2,5	0	1,5	2,5	2,5	0,1
Tomatenpüree (kg)	1,9	0	0	0	0	0	0
Gehackte Tomaten (kg)	50,3	52,0	40,2	44,0	60,4	50,3	6,3
Oregano frisch (g)	500	200	0	150	0	0	25
Zucker (g)	10	0	0	100	0	0	12

Zutat	Referenz (BB)	ChatGPT 4	ChatGPT 3.5	Perplexity	DeepSeek	Grok	Claude
Salz (Sauce, g)	151	500	0	200	402	1010	25
Knoblauch (g)	0	200	0	250	0	0	50

Tab.: Mengenvergleich ausgewählter Saucen-Zutaten – Referenz vs. KI-Plattformen. (Zutaten, die von allen KIs weggelassen wurden, wie Lorbeerblatt, sind hier nicht aufgeführt. **0** bedeutet, dass die Zutat nicht angegeben wurde.)

- **Zwiebel:** Das Referenzrezept enthält Zwiebel (umgerechnet ca. 18,9 kg für 503 Portionen). **Alle KI-Plattformen liessen die Zwiebel komplett weg.** Der Durchschnitt über die KIs ist somit 0 kg (Spannweite 0–18,9 kg gegenüber Referenz). Dies bedeutet, dass **keine KI die im Referenzrezept vorgesehene Zwiebel in der Sauce berücksichtigt hat.**
- **Olivenöl (für die Sauce):** Betty Bossi verwendet ~1,9 l Olivenöl zum Andünsten der Sauce. Die meisten KIs liegen hier auf ähnlichem Niveau oder etwas höher: ChatGPT 4, DeepSeek und Grok nennen jeweils ~2,5 l (etwa +30%). Perplexity gibt 1,5 l an (ca. 20% weniger als Referenz). **ChatGPT 3.5 verzichtet jedoch völlig auf Öl in der Sauce**, während Claude mit 0,1 l nur eine minimale Menge angibt. Im Mittel wurden ~1,8 l vorgeschlagen (nahe am Referenzwert), aber die Standardabweichung (~0,9 l) zeigt, dass insbesondere eine Plattform (GPT 3.5) deutlich abweicht. **Die meisten KI-Angaben stimmen moderat mit der Referenz überein, doch einzelne KIs (GPT 3.5, Claude) vergessen oder vernachlässigen das Öl in der Sauce.**
- **Tomatenpüree (Tomatenmark):** Laut Referenz ~1,9 kg Tomatenpüree. **Keine der KI-Plattformen hat Tomatenpüree in der Sauce vorgesehen** (alle Werte 0). Hier besteht vollständige Abweichung: Der Durchschnitt der KI-Mengen ist 0 kg gegenüber ~1,9 kg Referenz – **somit wurde das Tomatenmark von allen KIs ausgelassen.** Dies könnte bedeuten, dass die KIs entweder das Rezept vereinfachen oder das Konzentrat für entbehrlich halten.
- **Gehackte Tomaten (Dosentomaten):** Die Referenz verwendet ~50,3 kg Dosentomaten (entsprechend 100 g pro Pizza). Die KI-Werte liegen im **Schnitt bei ~42,2 kg**, also deutlich darunter (ca. 84% des Referenzwerts). Allerdings variiert dies stark: **DeepSeek schlägt ~60,4 kg vor (+20% mehr als Referenz)**,

während ChatGPT 3.5 nur ~40,2 kg verwendet (rund 20% weniger). ChatGPT 4 (52,0 kg) und Grok (50,3 kg) treffen die Referenz ziemlich genau. Perplexity liegt mit ~44,0 kg etwas darunter (–12%). **Claude sticht negativ heraus: lediglich ~6,3 kg Tomaten**, was nur etwa 12% der erforderlichen Menge ist. Die Spannweite reicht von 6,3 kg bis 60,4 kg – ein enormer Unterschied. Abgesehen von Claude bewegen sich die KI-Plattformen zwischen ~80–120% der Referenzmenge. Die Standardabweichung (≈17,3 kg) ist relativ hoch, was vor allem an Claudes Ausreisser liegt. **Insgesamt tendieren die meisten KI-Rezepte zu etwas weniger Tomate als das Referenzrezept – mit einer eklatanten Ausnahme (Claude).**

- **Oregano (frisch) in der Sauce:** Betty Bossi verwendet frischen Oregano (0,5 Zweiglein pro Pizza, gesamt ~500 g). Die Hälfte der KIs (ChatGPT 3.5, DeepSeek, Grok) **verwenden keinen frischen Oregano** in der Sauce. Die anderen Hälfte schon: ChatGPT 4 gibt ~200 g an, Perplexity 150 g, Claude 25 g. Im Mittel sind das ~62 g, was deutlich unter dem Referenzwert liegt – allerdings aufgrund der vielen Nulldaten. Betrachtet man nur die drei KIs, die Oregano verwenden, liegen deren Mengen bei 25–200 g (Referenz: 500 g). **Keiner der KI-Vorschläge erreicht die im Referenzrezept vorgesehene Menge an frischem Oregano; einige verzichten ganz darauf oder fügen stattdessen getrockneten Oregano erst als Belag hinzu (siehe unten).**
- **Lorbeerblatt:** Die Referenz enthält auch Lorbeer (ca. 125,8 Blätter gesamt). **Dieses Detail wurde von allen KI-Plattformen übersehen** – keine hat Lorbeer in der Sauce angegeben. (In der Tabelle oben ist Lorbeer nicht aufgeführt, da überall „kein“.) Damit fehlt diese Zutat bei 100% der KI-Rezepte.
- **Zucker:** Um die Säure zu balancieren, enthält die Referenz eine kleine Menge Zucker (ca. 10 g auf die ganze Sauce verteilt). Die Mehrheit der KIs **verzichtet auf Zucker**: ChatGPT 4, GPT 3.5, DeepSeek und Grok gaben **keinen Zucker** an. Claude nannte 12 g (in der Grössenordnung der Referenz, +20%). Perplexity sticht mit **100 g Zucker** hervor – das ist das **Zehnfache** der Referenzmenge. Durchschnittlich wurden ~18,7 g Zucker verwendet (bei einer Standardabweichung von ~36,6 g, was die Ausreisser verdeutlicht). **Fazit:** Vier KIs lassen den Zucker komplett weg (möglicherweise ignoriert oder bewusst als unnötig erachtet), eine (Claude) trifft den Referenzwert ungefähr, und eine (Perplexity) verwendet *sehr* viel Zucker.

- **Salz (für die Sauce):** Die Referenz würzt die Sauce zusätzlich mit ~150 g Salz. Hier zeigen sich erhebliche Unterschiede in der **Würzung der Sauce** durch die KIs:
    - **Zwei Plattformen (ChatGPT 3.5 und DeepSeek)** scheinen **gar kein Salz in der Sauce** vorgesehen zu haben (GPT 3.5 explizit „kein“, DeepSeek ebenfalls nicht separat angegeben – vermutlich wurde dort vergessen, die Sauce zu salzen).
    - Die anderen vier KIs würzen schon, allerdings sehr unterschiedlich stark: ChatGPT 4 schlägt 500 g vor (über 3× so viel wie Referenz), Perplexity 200 g (+33%), DeepSeek hingegen 402 g (wenn angegeben, +168%) und **Grok sogar 1010 g**, also über **1 kg Salz nur für die Sauce** (das ist >6× der Referenzmenge und würde die Sauce extrem versalzen machen). Claude gibt lediglich 25 g an (nur 1/6 der Referenz – die Sauce bliebe eher fade).
    - Diese Diskrepanz spiegelt sich auch im Durchschnitt (~349 g) und der enormen Spannweite (0 bis 1010 g) wider. **Mehrere KI-Plattformen haben entweder vergessen, die Sauce abzuschmecken (0 g Salz) oder massiv überwürzt**. Besonders Grok’s Wert ist eine krasse Abweichung. Offensichtlich besteht hier eine hohe Inkonsistenz darin, wie (und ob) die KI-Rezepte die Sauce salzen.
- **Knoblauch:** Interessanterweise enthält das Betty Bossi Rezept **keinen Knoblauch** – ein klassisches Pizza Napoli Rezept setzt offenbar eher auf Zwiebeln als Basis. Einige KI-Plattformen haben jedoch Knoblauch ergänzt: ChatGPT 4 nennt 200 g, Perplexity 250 g, Claude 50 g (entspricht z.B. grob 10–50 Knoblauchzehen, je nach Grösse). Die übrigen drei (GPT 3.5, DeepSeek, Grok) fügen **keinen Knoblauch** hinzu. Da die Referenz hier 0 hat, kann man keine Abweichung berechnen; qualitativ ist aber festzuhalten: **Die KIs sind uneins, ob Knoblauch in eine Pizza-Napoli-Sauce gehört** – die Hälfte tut es (in moderaten Mengen), die andere Hälfte nicht.

Belag-Zutaten

Schliesslich der Vergleich der Zutaten für den Pizzabelag. Dazu gehören Käse und typische Napoli-Garnituren wie Sardellen, Kapern, Oliven, etc. (Basilikum und ähnliches Kräuter-Topping werden im nächsten Abschnitt separat betrachtet, da sie im Referenzrezept nicht vorkommen.)

Die folgende Tabelle zeigt die Mengen der wichtigsten Belag-Zutaten (alle Angaben für 503 Pizzen):

Zutat	Referenz (BB)	ChatGPT 4	ChatGPT 3.5	Perplexity	DeepSeek	Grok	Claude
Tomaten (frisch, kg)	25,2	0	0	0	0	0	0
Mozzarella (kg)	18,9	75,0	50,3	151,0	50,3	50,3	6,3
Sardellenfilets (kg)	25,2	9,0	7,6	16,0	7,6	10,1	2,5
Kapern (kg)	2,26	8,00	5,03	3,00	2,52	7,55	0,15
schwarze Oliven (kg)	12,6	0	0	11,0	0	0	0
Oregano getr. (kg)	0,75	0	0,25	0	0,50	0,50	0

Tab.: Mengenvergleich der Belag-Zutaten – Referenz vs. KI-Plattformen. (Angaben in kg; 0 = nicht verwendet.)

- **Frische Tomaten (Stücktomaten auf der Pizza):** Das Referenzrezept belegt die Pizza noch mit frischen Tomaten (0,5 Stk. pro Pizza, ~25 kg gesamt). **Keine der KI-Plattformen hat jedoch frische Tomatenscheiben o.Ä. als Belag eingeplant.** Alle Werte sind 0 kg gegenüber 25,2 kg Referenz. Offenbar wurde dieser Schritt von den KIs durchwegs ausgelassen – vermutlich fokussierten sie sich auf die Tomatensauce und andere typische Napoli-Zutaten (Anchovies, etc.), während Betty Bossi zusätzlich frische Tomaten auflegt.
- **Mozzarella:** Hier treten **sehr grosse Abweichungen** auf. Die Referenz (hochgerechnet) sieht ~18,9 kg Mozzarella vor – das entspricht nur ca. 37,5 g Käse pro Pizza, was relativ wenig ist. Die KI-Vorschläge liegen fast alle deutlich höher:

- **Perplexity gibt mit 151 kg eine extrem hohe Käsemenge** an – das sind ~300 g Mozzarella pro Pizza (etwa 8× so viel wie die Referenz). Dies ist mit Abstand der höchste Wert und wirkt im Vergleich sehr überdimensioniert.
- ChatGPT 4 schlägt 75,0 kg vor (~149 g pro Pizza), also rund **4-fache Referenzmenge**. ChatGPT 3.5, DeepSeek und Grok landen alle bei ~50,3 kg (exakt 100 g pro Pizza, etwa das **Dreifache** der Referenz). Interessanterweise stimmen diese drei Plattformen hier überein – möglicherweise haben sie ähnliche Faustregeln (100 g Käse pro Pizza ist ein gängiger Richtwert).
- Claude fällt erneut aus dem Rahmen mit nur **6,3 kg Mozzarella** – das sind ~12,5 g pro Pizza (kaum ein Biss Käse). Dieser Wert beträgt nur ~33% der Referenzmenge und ~1/20 der durchschnittlichen KI-Menge.

Im Durchschnitt verwendeten die KIs ~63,9 kg Käse, also **gut das Dreifache der Referenz**. Die Spannweite (6,3 kg bis 151 kg) ist enorm, und die Standardabweichung (~44 kg) ebenfalls – grösser als der Mittelwert. **Fazit:** Fast alle KIs bestücken die Pizza wesentlich käsereicher als das Referenzrezept (was eher spärlich Käse vorsieht, wohl um den Geschmack der Sardellen nicht zu überdecken). Perplexity schiesst dabei weit übers Ziel hinaus, während Claude auf der anderen Seite viel zu wenig Käse einplant.

- **Sardellenfilets (Anchovies):** Die Referenz verwendet reichlich Sardellen (~25,2 kg für 503 Pizzen, d.h. 50 g pro Pizza, ca. eine halbe kleine Dose). Hier tendieren die KI-Ergebnisse geschlossen zu **deutlich niedrigeren Mengen**:
  - Die Bandbreite reicht von **2,5 kg** (Claude) bis **16,0 kg** (Perplexity). Selbst der höchste KI-Wert (Perplexity, ~16 kg) entspricht nur etwa 64% der Referenzmenge.
  - ChatGPT 4 schlägt 9,0 kg vor (~18 g pro Pizza), ChatGPT 3.5 und DeepSeek je ~7,6 kg (15 g pro Pizza), Grok ~10,1 kg (20 g pro Pizza). Diese Mengen liegen im Bereich von **2–4 Sardellenfilets pro Pizza**, was realistischer erscheint als die ~50 g (vielleicht ~8 Filets) der Referenz.

Durchschnittlich nutzten die KIs ~8,78 kg Sardellen – **nur etwa 35% der Referenzvorgabe**. Die geringe Standardabweichung relativ zum Mittel (≈4 kg) zeigt, dass die meisten KI-Werte in ähnlicher Grössenordnung liegen. **Alle KI-Plattformen reduzierten die Sardellenmenge drastisch im Vergleich zum Betty-Bossi-Rezept.**

Dies wirkt plausibel, da 50 g Sardellen pro Pizza sehr salzig und geschmacksintensiv wären; offenbar haben die KI-Rezepte hier Zurückhaltung geübt.

- **Kapern:** Die Referenz setzt ca. 2,26 kg Kapern ein (≈4,5 g pro Pizza). Die KI-Vorschläge variieren stark:
    - ChatGPT 4 und Grok nennen **8,0 kg bzw. 7,55 kg** Kapern – etwa das **Drei- bis Vierfache** der Referenz (entspricht ~15 g Kapern pro Pizza, also z.B. ein gehäufter Esslöffel voll, sehr viel).
    - ChatGPT 3.5: 5,03 kg (~10 g/Pizza, ~Doppelte der Referenz). Perplexity: 3,00 kg (~6 g/Pizza, geringfügig über Referenz). DeepSeek: 2,52 kg (~5 g/Pizza, nahezu identisch mit Referenz). **Claude:** 0,15 kg, also **150 g gesamt** – das sind vernachlässigbare ~0,3 g pro Pizza (vielleicht eine einzelne Kaper auf jede dritte Pizza). Dieser Wert ist um den Faktor 15 niedriger als die Referenz.

Der KI-Durchschnitt (~4,37 kg) liegt knapp doppelt so hoch wie die Referenz, jedoch bei hoher Streuung (Std ~2,79 kg). **Einige KIs übertreiben die Kapernmenge deutlich (bis zu 3–4× Referenz), während Claude praktisch keine Kapern vorsieht.** Die restlichen bewegen sich zwischen +0% und +100% des Referenzwerts.

- **Schwarze Oliven:** Laut Referenz ~12,6 kg (ca. 25 g pro Pizza). **Nur eine KI (Perplexity) hat Oliven eingeplant** – mit ~11,0 kg kommt Perplexity dem Referenzwert sogar nahe (≈22 g/Pizza). Alle anderen fünf Plattformen **führen keinerlei Oliven** auf. Der Durchschnitt über alle KIs beträgt demnach ~1,83 kg (wobei diese Zahl wenig sinnvoll ist, da 5 von 6 Null sind). Wichtig ist: **Die grosse Mehrheit der KI-Plattformen hat die Oliven weggelassen**, obwohl sie im Originalrezept vorgesehen sind.
- **Oregano (getrocknet) als Belag:** Das Referenzrezept streut am Schluss getrockneten Oregano über die Pizza (0,5 EL pro Pizza, ~0,755 kg gesamt). Hier haben diejenigen KIs zugelangt, die *keinen* frischen Oregano in der Sauce verwendet hatten:
    - ChatGPT 3.5 gibt ~0,25 kg getrockneten Oregano an (ca. 0,5 g pro Pizza, ein üblicher Prise-Wert).
    - DeepSeek und Grok verwenden sehr viel: beide ~0,50 kg (etwa 1 g pro Pizza oder 1 TL – das ist kräftig, aber noch vertretbar).

- ChatGPT 4, Perplexity und Claude streuen **gar keinen** getrockneten Oregano (sie hatten dafür frischen in der Sauce bzw. Basilikum als Garnitur, siehe unten).

Der Schnitt liegt bei ~0,54 kg, was leicht über dem Referenzwert ist, doch mit hoher Varianz. **In Summe haben alle Plattformen Oregano entweder in frischer oder getrockneter Form eingesetzt** – keine liess das Würzkraut komplett weg. Allerdings nutzt etwa die Hälfte (z.B. GPT 4, Perplexity) den Oregano bereits in der Sauce (frisch) und verzichtet auf das finale Bestreuen, während die andere Hälfte (Grok, DeepSeek, GPT 3.5) exakt umgekehrt verfährt. Die Referenz wiederum tut beides (frisch in Sauce *und* getrocknet auf Pizza) – hierin stimmen die KI-Rezepte also **nur teilweise** überein (sie implementieren jeweils nur eine der beiden Methoden, nicht beide).

2. Vollständigkeit und Übereinstimmung der Zutaten

In diesem Abschnitt wird geprüft, **ob alle KI-Plattformen die nötigen Zutaten des Referenzrezepts abgedeckt haben**, und welche Zutaten eventuell fehlen oder hinzugefügt wurden. Dabei ergeben sich pro Plattform folgende Befunde:

- **ChatGPT 4 (OpenAI):** Enthält *fast* alle wichtigen Zutaten des Referenzrezepts. **Fehlend** (vergessen) sind: Zwiebel, frische Tomaten (als Belag), Lorbeerblatt, Tomatenpüree, getrockneter Oregano (als Abschluss), Zucker sowie Pfeffer. **Zusätzlich** (nicht im Betty Bossi Rezept) verwendet ChatGPT 4: Basilikum (als frisches Kraut zum Garnieren), etwas extra Öl zum Beträufeln der Pizza und Knoblauch in der Sauce. Insgesamt ist ChatGPT 4 recht vollständig – alle essentiellen Kernzutaten (Mehl, Hefe, Salz, Wasser, Olivenöl, Tomaten, Mozzarella, Sardellen, Kapern) sind vorhanden. Kleinere Zutaten wie Zucker oder Lorbeer wurden jedoch ausgelassen, und Pfeffer zum Abschmecken fehlt ebenfalls.
- **ChatGPT 3.5:** Lässt einige Punkte aus. **Fehlende Zutaten:** Zwiebel, frische Tomaten, Lorbeer, Tomatenpüree, *frischer* Oregano (in der Sauce), Zucker, Pfeffer **und sogar Olivenöl** zum Kochen (weder im Teig noch in der Sauce – nur als Finish-Öl genutzt). ChatGPT 3.5 hat also keinen Zwiebel-Knoblauch-Ansatz (verzichtet sowohl auf Zwiebel als auch Knoblauch), kein Zucker, kein Lorbeer – die Sauce besteht im Prinzip nur aus Tomaten, Salz und getrocknetem Oregano. **Zusätzlich** verwendet auch GPT 3.5 ein wenig Öl zum Beträufeln nach dem Backen (was im Referenzrezept nicht explizit vorgesehen ist). Es gibt keine weiteren Extra-Zutaten. Insgesamt hat ChatGPT 3.5 wichtige

Geschmackszutaten (Zwiebel, Knoblauch, Zucker, Pfeffer) weggelassen und auch das Öl in Teig/Sauce vergessen, was die Vollständigkeit spürbar einschränkt.

- **Perplexity:** Deckt viele Aspekte ab, hat aber auch Lücken. **Fehlend gegenüber Referenz:** Zwiebel, frische Tomaten, Lorbeerblatt, Tomatenpüree, *getrockneter* Oregano (zum Schluss). **Zusätzliche Zutaten:** Basilikum (als Garnitur), Knoblauch in der Sauce, extra Olivenöl zum Beträufeln und interessanterweise **Pfeffer „nach Geschmack“** (als einzige KI hat Perplexity explizit Pfeffer zum Abschmecken genannt). Zudem hat Perplexity – anders als die meisten – **Oliven** aufgelistet (nahezu in Referenzmenge). Von den Referenzzutaten, die viele KIs vernachlässigten, hat Perplexity also Zucker (ja, sogar im Übermass), Pfeffer und Oliven berücksichtigt, aber z.B. Zwiebeln und Lorbeer weiterhin nicht. Insgesamt erscheint Perplexity vergleichsweise **komplett**, da es neben den Grundzutaten auch an Würzungen wie Pfeffer, Knoblauch und Zucker gedacht hat – im Referenzvergleich fehlen hauptsächlich die Zwiebeln und das Tomatenmark.
- **DeepSeek:** Auch DeepSeek hat einen Grossteil der Basiszutaten drin. **Fehlend** sind allerdings: Zwiebel, frische Tomaten, Lorbeerblatt, Tomatenpüree, *frischer* Oregano (in der Sauce), Zucker und Pfeffer. **Zusätzlich** hinzugefügt wurden: frisches Basilikum (allerdings unüblicherweise in die Sauce integriert, als „503 Stängel“), sowie etwas Olivenöl zum Beträufeln nach dem Backen. Knoblauch taucht in DeepSeeks Zutatenliste nicht auf (im Gegensatz zu manch anderer KI). Insgesamt hat DeepSeek die typischen Napoli-Zutaten (Anchovies, Kapern, Olivenöl, etc.) beisammen, verzichtet aber – wie viele – auf Zwiebel, Zucker und Lorbeer und bietet keine explizite Pfeffer- oder Knoblauchwürze. Das Rezept ist somit **nicht vollständig identisch** mit dem Referenzsatz an Zutaten, aber alle unverzichtbaren Belagskomponenten sind enthalten.
- **Grok:** Bei Grok fehlen einige der gleichen Elemente. **Fehlend** laut Referenz: Zwiebel, frische Tomaten, Lorbeer, Tomatenpüree, *frischer* Oregano, Zucker, Pfeffer. **Zusätzlich** taucht bei Grok einzig Basilikum als extra Zutat auf. (Grok hat Basilikum interessanterweise doppelt in Sauce und Belag erwähnt – vermutlich als frisches Kraut – was im Referenzrezept nicht vorkommt.) Ansonsten hat Grok keine weiteren Extras wie Knoblauch oder Olivenöl-Drizzle angegeben (es belässt es bei den Standardzutaten). Das heisst, **Grok hat im Grunde dieselbe Lücke** wie GPT 3.5 und DeepSeek: es vergisst die feinen Würzzutaten (Zwiebel, Knoblauch, Zucker, Pfeffer, Lorbeer) und setzt nur auf das Nötigste. Damit ist

Groks Zutatenliste zwar fokussiert, aber in einigen Punkten **unvollständig im Vergleich zur Referenz**.

- **Claude:** Das Rezept von Claude zeigt **erhebliche Auslassungen**. **Fehlend** sind: Zwiebel, frische Tomaten, Lorbeer, Tomatenpüree, *getrockneter* Oregano (am Schluss) **sowie Pfeffer**. Interessant ist, dass Claude im Gegensatz zu vielen anderen **Zucker** nicht vergessen hat (12 g angegeben) und auch **Knoblauch** (50 g) und Basilikum (250 g) als Zusatz enthält. Zudem nennt Claude eine kleine Menge Öl zum Beträufeln. **Zusätzlich** im Vergleich zur Referenz sind also: Basilikum, Knoblauch und Finish-Öl (ähnlich wie bei ChatGPT 4 und Perplexity). Trotz dieser Extras weist Claude gravierende Lücken auf: wichtige Basiszutaten wie ausreichend Mehl, Wasser oder Salz (siehe Mengenvergleich) stimmen nicht, was letztlich eine grössere Abweichung darstellt als das Fehlen einzelner Gewürze. Insgesamt ist Claudes Zutatenliste nicht umfassend – sie hat zwar an manche Würzmittel gedacht, dafür aber zentrale Mengen falsch berechnet (siehe nächster Abschnitt), was das Rezept als Ganzes in Frage stellt.

Zusammengefasst ergibt sich folgendes Bild der **Vollständigkeit**:

- **Alle** KI-Plattformen haben gewisse Zutaten des Betty Bossi Rezepts **ausgelassen**. Besonders **Zwiebeln, Lorbeerblätter und Tomatenpüree** fehlen in *jedem* KI-Rezept – kein einziger KI-Vorschlag enthält diese vom Referenzrezept vorgesehenen Komponenten. Ebenso haben fast alle KIs auf **frische Tomaten als Belag** verzichtet (eine Zutat, die evtl. als verzichtbar angesehen wurde).
- **Häufig fehlende** Zutaten (bei mindestens der Hälfte der KIs): Zucker (4 von 6 KIs weggelassen), Pfeffer (5 von 6 weggelassen oder nicht quantifiziert), Oliven (5 von 6 weggelassen) und Knoblauch (3 von 6 weggelassen, 3 haben ihn hinzugefügt). **Oregano** wurde von allen benutzt, aber auf unterschiedliche Weise – die einen frisch (Sauce), die anderen getrocknet (Belag) – sodass jede einzelne Variante (frisch vs. getrocknet) im Vergleich zur Referenz unvollständig ist (Referenz nutzt beides).
- **Zusätzliche Zutaten**, die im Betty Bossi Rezept nicht vorkommen, wurden von mehreren KIs eingeführt: Allen voran **Basilikum** (bei 5 von 6 KI-Rezepten hinzugefügt) und **Knoblauch** (bei 3 von 6 hinzugefügt). Auch das **Beträufeln mit Öl** nach dem Backen taucht bei 5 von 6 KI-Vorschlägen auf, obwohl es im Referenzrezept nicht explizit erwähnt wird. Diese Zusätze sind kulinarisch

plausibel (Basilikum und Knoblauch sind in vielen Pizza-Rezepten üblich, ebenso ein Schuss gutes Öl zum Schluss), zeigen aber, dass die KIs teils von der Vorlage abweichen und eigene Akzente setzen.

Im Ergebnis hat **keine** der KI-Plattformen das Zutaten-Set der Referenz 1:1 getroffen. Dennoch decken alle KIs die grundlegenden Erfordernisse (Teigzutaten, Tomaten, Käse, Anchovies, Kapern) ab. Die Unterschiede liegen vor allem in den **sekundären Zutaten** (Würzmittel, Kräuter, Zusatzöl) – mal fehlen diese komplett, mal werden sie anders oder in abweichender Menge eingesetzt.

3. Qualitative Auffälligkeiten

Bei der Auswertung der Rezepte zeigen sich einige qualitative Besonderheiten – sei es ungewöhnliche Zutaten, extrem abweichende Mengen oder auffällige Formulierungen. Hier einige hervorstechende Punkte:

- **Universelles Weglassen von Zwiebel & Lorbeer:** Auffällig ist, dass *alle* KI-Plattformen sowohl auf **Zwiebeln** in der Sauce als auch auf das **Lorbeerblatt** verzichtet haben. Diese beiden Zutaten scheinen von keinem Modell als nötig erachtet worden zu sein, obwohl sie im Referenzrezept vorgesehen sind. Das deutet darauf hin, dass KI-Modelle entweder selten Zwiebeln/Lorbeer für Pizza-Sauce empfehlen (vielleicht weil viele Standardrezepte darauf verzichten) oder diese Details bei der Skalierungsaufgabe übersehen haben.
- **Tomatenmark (Tomatenpüree) ignoriert:** Ebenfalls interessant – kein KI-Rezept enthielt Tomatenmark. Das Referenzrezept nutzt eine kleine Menge, vermutlich zur Geschmacksvertiefung. Die KIs haben diesen Schritt sämtlich ausgelassen. Möglicherweise kennen/wählen viele KI-Modelle eher einfachere Saucenrezepte (nur mit passierten oder stückigen Tomaten) und liessen daher das Mark weg.
- **Gewürz- und Zuckerzugabe uneinheitlich:** Die Rolle von **Zucker und Pfeffer** im Saucenabschmecken wurde sehr uneinheitlich gehandhabt. Mehrere KIs (z.B. GPT 4, GPT 3.5, DeepSeek, Grok) haben **weder Zucker noch Pfeffer** für die Sauce erwähnt – was bedeutet, dass sie die Säure und Würze der Tomaten nicht explizit ausbalancieren. Perplexity hingegen gab relativ viel Zucker („100 g, nach Geschmack“) hinzu und als einzige Plattform „Pfeffer nach Geschmack“. Claude fügte wenig Zucker hinzu, aber keinen Pfeffer. Diese Unterschiede zeigen, dass

die KI-Plattformen beim Thema **Abschmecken** der Sauce sehr unterschiedlich vorgehen: Manche scheinen davon auszugehen, dass ein Pizza-Rezept keinen Zucker/Pfeffer braucht (oder vergessen es schlicht), während andere es erwähnen. Besonders Perplexitys Zuckerangabe (100 g) ist **ungewöhnlich hoch** – das könnte ein Missverständnis („nach Bedarf“ bei grossen Mengen) sein.

- **Käse-Extremwerte:** Die Mozzarella-Menge von **Perplexity (151 kg)** ist ein extremer Ausreisser. Sie würde ~300 g Käse pro Pizza bedeuten – was für eine einzelne Pizza übertrieben viel ist (das entspricht fast 3 Packungen Mozzarella pro Pizza). Dieses Übermass sticht deutlich hervor und wäre in der Praxis wohl **übertrieben**. Im Gegensatz dazu steht **Claude mit nur 6,25 kg** (rund 12 g pro Pizza, also fast kein Käse) am anderen Ende der Skala. Beide sind klare Ausreisser, die auf Rechenfehler oder falsche Annahmen hindeuten. Die restlichen KIs pendeln um ~100–150 g Käse/Pizza, was realistisch erscheint. Dieses Beispiel illustriert, wie **uneinheitlich die KI-Antworten skalieren**: Einige übersetzen „für 503 Personen“ fast mechanisch (und riskieren krasse Werte), während andere anscheinend vergessen, Teile des Rezepts mit zu skalieren (Claude).
- **Hefe-Fehlmengen bei GPT 3.5 & Perplexity:** Ein besonders kritischer Punkt ist die **viel zu geringe Hefemenge** bei zwei Plattformen: ChatGPT 3.5 (40 g) und Perplexity (50 g). Diese Mengen wären für ~500 kg Teig (so viel Mehl wurde ja angegeben) absolut unzureichend – der Teig würde kaum aufgehen. Es wirkt, als hätten diese Modelle die Hefe entweder *nicht skaliert* (z.B. eine Standardmenge für ein normales Rezept übernommen) oder bewusst minimal gehalten (was aber technisch nicht passt). Dies ist qualitativ gravierend, da es das Gelingen des Teiges infrage stellt. Im Gegensatz dazu haben DeepSeek und Grok eher zu viel Hefe angesetzt (was weniger problematisch ist als zu wenig, aber geschmacklich unsauber). ChatGPT 4 hat die Hefe deutlich reduziert im Vergleich zur linearen Skalierung (40% der Referenzmenge) – das könnte bewusst sein (längere Gehzeit, weniger Hefe nötig) oder ein Versehen. Jedenfalls zeigen diese Extreme, dass **KIs bei Skalierungen leicht Fehler machen**: Entweder vergessen sie die Proportionen (wie GPT 3.5/Perplexity hier) oder sie skalieren streng linear, obwohl man Hefe oft *nicht* linear skaliert für grosse Mengen (eine moderate Reduktion ist plausibel, wie GPT 4 tat).
- **Salz und Würze inkonsistent:** Die **Salzangaben** schwanken stark zwischen den KI-Rezepten. Besonders grotesk ist **Groks Vorschlag, über 1 kg Salz in die Sauce zu geben** – das wäre ungeniessbar. Möglicherweise entstand das,

weil Grok bereits den Teig stark gesalzen hat und dann die gleiche Salzdosierung irrtümlich nochmals für die Sauce ansetzte. Umgekehrt hat GPT 3.5 scheinbar vergessen, die Sauce überhaupt zu salzen. Solche **fehlenden oder doppelt gemoppelten Würzungen** sind qualitativ auffällig: Sie deuten auf fehlende Feinkontrolle in den KI-Antworten hin. Ebenfalls bemerkenswert: Perplexity's Formulierung „Pfeffer nach Geschmack" zeigt an, dass dort keine genaue Menge genannt wurde – diese Freitextangabe ist qualitativ anders als die streng numerischen Angaben der anderen. Es erschwert die direkte Mengenvergleichbarkeit (in unserer Auswertung wurde Pfeffer bei Perplexity effektiv als „0" gezählt, obwohl es qualitativ vorhanden war).

- **Basilikum fast überall, obwohl Referenz „kein":** Fünf der sechs KI-Rezepte fügen **Basilikum** hinzu (frische Blätter), obwohl es laut Betty Bossi nicht vorgesehen ist. Basilikum ist eine klassische italienische Zutat und vermutlich haben die KI-Modelle aus allgemeinen Rezeptkenntnissen „gelernt", dass Basilikum oft passt – daher taucht es fast überall auf, teils als Belag (z.B. ChatGPT 4: 1 kg; Perplexity: 600 g; Claude: 250 g frische Basilikumblätter) oder sogar in der Sauce (DeepSeek: ein Basilikumstängel pro Portion). Das ist eine qualitative Abweichung im Ablauf: **DeepSeek gibt Basilikum in die Sauce statt als Topping** – ein etwas ungewöhnlicher Ansatz (frisches Basilikum verliert beim Mitkochen an Aroma, klassischer wäre es, die Sauce mit Basilikum zu ziehen und die Stängel wieder zu entfernen – DeepSeek erwähnte aber nur „503 Stängel", unklar ob zum Mitkochen oder Dekorieren). Grok hat Basilikum interessanterweise doppelt aufgeführt (in Sauce und Belag jeweils 2,52 kg) – das wirkt wie ein Datenfehler oder die Interpretation, man solle sehr grosszügig Basilikum verwenden. Qualitativ sind Basilikumbätter als Garnitur natürlich kein Fehler, aber **die Referenz hielt sich daran, wohl weil Oregano und Kapern/Anchovies geschmacklich dominieren sollten**. Dass fast alle KIs Basilikum hinzufügen, zeigt eine gewisse Tendenz, Rezepte mit typischen italienischen Zutaten anzureichern, selbst wenn das konkrete Originalrezept diese nicht vorsieht.
- **Zusatz von Knoblauch in einigen KI-Rezepten:** Ähnlich wie Basilikum haben einige KIs **Knoblauch** ergänzt, obwohl das Referenzrezept keinen enthält. Perplexity (250 g), ChatGPT 4 (200 g) und Claude (50 g) gaben Knoblauchzehen in die Sauce. Das ist qualitativ nachvollziehbar – viele Pizzasaucen enthalten entweder Zwiebel oder Knoblauch (oder beides). Dass Betty Bossi hier auf Zwiebel setzt und die KI-Modelle teils auf Knoblauch wechseln, ist eine

interessante Abweichung. Allerdings haben ChatGPT 3.5, DeepSeek und Grok *weder* Zwiebel noch Knoblauch – ihre Saucen basieren rein auf Tomate und ggf. Oregano, was geschmacklich etwas flach sein könnte. Insgesamt spiegeln diese Unterschiede, wie die KI sich typische Aromaten aussuchen: Manche **ersetzen die Zwiebel durch Knoblauch**, andere lassen beides weg (vermutlich nicht ideal), keiner verwendet (wie die Referenz) die Kombination aus Zwiebel *ohne* Knoblauch, ausser GPT 4 (das aber auch Knoblauch hinzugab, also Zwiebel *mit* Knoblauch wäre keins; tatsächlich hat GPT 4 ebenfalls keine Zwiebel). Qualitativ fällt hier besonders negativ auf, dass **kein KI-Rezept den originalen Zwiebelfond nachahmt** – wer sich streng ans KI-Rezept hält, würde eine anders schmeckende Sauce erhalten.

- **Fehlende oder unpräzise Mengeneinheiten in KI-Antworten:** Bei der Konsolidierung fiel auf, dass einige KI-Angaben in wenig präzisen Begriffen gemacht wurden, was qualitativ auffällt. Beispielsweise schrieb Perplexity **„Pfeffer je nach Geschmack“** und **„Olivenöl zum Beträufeln“** – also ohne konkrete Mengenangabe. Für unsere Analyse mussten solche Fälle als „kein numerischer Wert“ behandelt werden. In der Praxis zeigen solche Formulierungen aber eine **Unschärfe**: Die KI gibt dem Benutzer Freiraum („nach Geschmack“), was einerseits realistisch sein mag, andererseits aber das genaue Befolgen erschwert. Im Gegensatz dazu gaben andere KIs (z.B. GPT 4) für alles exakte Mengen, selbst für „1 Prise“ etc. Hier sieht man Unterschiede im Stil: Einige KIs neigen zu vagen Angaben bei grossen Mengen („abschmecken“), während andere streng skalieren.
- **Inkonsequente Abschnittseinteilung:** Eine kleinere Auffälligkeit ist die manchmal unklare Zuordnung von Zutaten zu Rezeptabschnitten. Zum Beispiel hat **Grok Basilikum sowohl unter „Sauce“ als auch unter „Belag“ gelistet** mit identischer Menge – was unwahrscheinlich ist (vermutlich ein Fehler oder Dopplung in der Antwort). **DeepSeek listete Basilikum unter Sauce (Stängel) statt als Belag**. Solche Abweichungen in der **Struktur** deuten darauf hin, dass KIs Rezepte nicht immer streng nach „Teig/Sauce/Belag“ gliedern oder eventuell Schritte vermischen. Für den Nutzer kann das irritierend sein – z.B. Basilikum sollte typischerweise am Schluss frisch auf die Pizza, nicht in die kochende Sauce, wie DeepSeek es impliziert.

Zusammengefasst sind die qualitativen Auffälligkeiten vor allem: **vergessene Zutaten (Zwiebel, Lorbeer, Tomatenmark), extreme Fehlskalierungen (Hefe, Salz, Käse), sowie variierende Rezeptinterpretationen** (z.B. Basilikum/Knoblauch-Einsatz, Würzung „nach Geschmack“). Diese Unterschiede zeigen, dass die KI-Plattformen nicht einheitlich „logisch“ skaliert oder an jedes Detail gedacht haben. Einige Fehler – etwa Claudes generelle Mengen – fallen so stark aus, dass sie das Rezept unbrauchbar machen würden (z.B. viel zu wenig Mehl, Wasser, Salz, etc.), während andere Abweichungen eher geschmacklich oder stilistisch ins Gewicht fallen (z.B. kein Zucker, dafür Basilikum hinzugefügt – was die Pizza nicht scheitern lässt, aber anders macht als vorgesehen).

4. Bewertung der KI-Plattformen

Abschliessend sollen die KI-Plattformen hinsichtlich **Genauigkeit**, **Vollständigkeit** und **Plausibilität** ihrer Rezeptvorschläge verglichen werden. Dabei zeigt sich ein differenziertes Bild:

- **ChatGPT 4 (OpenAI):** Insgesamt der **stimmigste und vollständigste Vorschlag**. ChatGPT 4 lieferte für alle Hauptzutaten plausible Mengen (Mehl, Wasser, etc. lagen nahe am Referenzwert) und vermied krasse Ausrutscher. Positiv fällt auf, dass alle wichtigen Belagzutaten vorhanden sind und die Teig- und Saucenmengen in einem vernünftigen Rahmen. Die Reduktion der Hefe auf 500 g (statt ~1,3 kg) könnte sogar als **plausible Anpassung** interpretiert werden (z.B. längere Gehzeit mit weniger Hefe). Zwar hat GPT 4 einige Kleinigkeiten vergessen (Zwiebel, Lorbeer, Zucker) und etwas viel Kapern verwendet, aber diese Abweichungen sind nicht kritisch. In Summe überzeugt ChatGPT 4 mit **hoher Genauigkeit** (kleinste durchschnittliche Abweichung zur Referenz bei den Kernzutaten), **guter Vollständigkeit** (fast alle relevanten Zutaten enthalten) und hoher **Plausibilität** im Sinne eines kochbaren Rezeptes. – **Platz 1** in der Gesamtbewertung.
- **DeepSeek:** DeepSeek lieferte ebenfalls überwiegend vernünftige Werte. Mehl und Wasser stimmten nahezu exakt mit der Referenz überein, und die meisten Mengen waren im grünen Bereich. Lediglich bei Salz (etwas hoch) und Hefe (etwas hoch, aber nicht völlig unplausibel) wich DeepSeek stärker ab. Vollständigkeit: Wie die meisten vergass auch DeepSeek Zwiebel, Lorbeer, Zucker und Pfeffer, und setzte Basilikum an eine ungewohnte Stelle (in die Sauce). Trotzdem waren alle Schlüsselingredienzen (Teigzutaten, Tomaten,

Käse, Sardellen, Kapern, Olivenöl) vorhanden. Plausibilität: Bis auf die versalzene Sauce (kein separates Salz angegeben – evtl. schlicht vergessen) klingt das Rezept brauchbar. Besonders positiv: Keine krassen Ausreisser wie bei manch anderer KI. Daher ist **DeepSeek** in Summe **sehr ordentlich** und knapp hinter ChatGPT 4 einzuordnen – leichte Abzüge für Auslassungen bei den Gewürzen. – **Platz 2**.

- **Perplexity:** Perplexity zeigte ein gemischtes Bild. Positiv ist die **Vollständigkeit**: Perplexity hat als einzige Plattform Pfeffer und Oliven berücksichtigt und auch an Knoblauch gedacht. Das Rezept wirkt damit würziger und relativ komplett. Allerdings gibt es zwei grosse Probleme in der **Genauigkeit** der Mengen: die **viel zu geringe Hefe** (50 g, was ein klarer Fehler ist) und die **riesige Käsemenge** (151 kg). Diese beiden Ausrutscher unterminieren die Plausibilität stark – der Teig würde kaum aufgehen und wäre mit Käse überfrachtet. Auch 100 g Zucker sind grenzwertig hoch. Abgesehen davon sind Perplexitys Mengen (Mehl, Wasser, Salz, Tomaten, Anchovies, Kapern) moderat und plausibel, oft etwas unter dem Referenzwert, aber im Rahmen. **Plausibilität:** Würde man Perplexitys Rezept kochen, müsste man definitiv die Hefe drastisch erhöhen und den Käse reduzieren, sonst scheitert es. Trotzdem sind die meisten Zutaten vorhanden und richtig, was Perplexity leicht vor die weniger vollständigen Rezepte hebt. Insgesamt ein **würziges, aber teils fehlerhaft skaliertes Rezept** – daher **Platz 3** (mittelmässig, mit gleichermassen Stärken und Schwächen).
- **Grok:** Grok lieferte solide Basisangaben (Mehl, Tomaten etc. ok), hat aber einige deutliche **Schwächen**: Das Weglassen sämtlicher Würzkomponenten (Zwiebel, Knoblauch, Pfeffer, Zucker) macht das Rezept recht eindimensional. Zudem hat Grok extreme **Salzüberschüsse** produziert (Teig und Sauce zusammen über 3 kg Salz – viel zu viel). Positiv ist, dass die meisten wichtigen Zutaten (bis auf Oliven) vorhanden sind und in vernünftiger Menge (Hefe war hoch, aber Teig hätte zumindest getrieben; Käsemenge mit 50 kg okay). Basilikum doppelt auszuweisen wirkt wie ein Fehler, könnte aber zu viel des Guten sein. **Plausibilität:** Durch den Salzfehler wäre Groks Pizza vermutlich versalzen. Auch die fehlenden Aromen (kein Zwiebel/Knoblauch) würden Geschmack kosten. Hier müsste ein Koch das Rezept nachbessern (weniger Salz, etwas Knoblauch zufügen). Insgesamt rangiert **Grok** im unteren Mittelfeld – **Platz 4** –, da zwar die Mengen grösstenteils skaliert wurden, aber einige grobe Patzer (Salz) und Auslassungen die Qualität mindern.

- **ChatGPT 3.5:** Dieses Rezept ist zwar ähnlich minimalistisch wie Grok, hat aber einen kritischen Fehler: die **Hefemenge** (40 g) ist so gering, dass der Teig praktisch nicht aufgeht. Auch fehlt sämtliches Öl im Teig und in der Sauce (was zu Teigtextur- und Geschmackseinbussen führt). Positiv: Die Kernzutaten Mehl/Wasser stimmen einigermaßen, und 100 g Käse pro Pizza sind vernünftig. Anchovies/Kapern hat GPT 3.5 moderat angesetzt. **Vollständigkeit** ist geringer als bei anderen – kein Zwiebel, kein Knoblauch, kein Zucker, kein Pfeffer, keine Oliven. Das ergibt ein eher fades Rezept. **Plausibilität:** Wegen des Hefe- und Öl-Fehlers ist die Machbarkeit eingeschränkt; der Teig könnte misslingen oder zumindest qualitativ schlecht werden. GPT 3.5 zeigt damit deutliche Schwächen sowohl in Genauigkeit (skaliert Hefe/Öl nicht korrekt) als auch in Vollständigkeit (vergisst mehrere Zutaten). Deshalb landet **ChatGPT 3.5** auf **Platz 5** – noch vor Claude, aber klar hinter den übrigen.
- **Claude:** Leider lieferte Claude ein Rezept, das in vieler Hinsicht **ungenügend** ist. Zwar hat Claude an einige Würzzutaten gedacht (Knoblauch, Basilikum, etwas Zucker), aber die **Hauptmengengerüste** für Teig und Sauce sind durchweg falsch skaliert – teils um den Faktor 1:10 zu niedrig. 12,5 kg Mehl für 503 Pizzen (statt ~62,5 kg) sind ein eklatanter Fehler; ähnlich drastisch unterschätzt wurden Wasser, Salz, Tomaten und Käse. Diese fundamentalen Abweichungen machen das Rezept praktisch unbrauchbar, da Mengen vorne und hinten nicht stimmen. Zwar ist Claude's Zutatenliste an sich relativ umfangreich, aber wenn die Mengen nicht passen, hilft das wenig. **Genauigkeit:** katastrophal (grösste Durchschnittsabweichung von der Referenz unter allen KIs). **Vollständigkeit:** mittel (einige Gewürze ja, anderes nein). **Plausibilität:** sehr gering – würde man nach Claudes Angaben kochen, bekäme man z.B. aus 12,5 kg Mehl niemals 503 Pizzen. Entsprechend muss **Claude als schlechtester Vorschlag** bewertet werden – **Platz 6** (letzter). Hier haben offenbar grobe Rechenfehler stattgefunden.

Zur Verdeutlichung eine **Rangliste der Plattformen** in Hinblick auf die Gesamtleistung:

1. **ChatGPT 4 (OpenAI)** – Präziseste Mengen, kaum grobe Fehler, Zutaten weitgehend vollständig.
2. **DeepSeek** – Ebenfalls sehr stimmig, mit minimalen Mängeln (fehlende Nebenzutaten).

3. **Perplexity** – Inhaltsreich (viele Zutaten), aber einige gravierende Skalierungsfehler (Hefe, Käse) drücken die Bewertung.
4. **Grok (OpenAI)** – Grundsätzlich ok, doch Gewürzarmut und Würzfehler (Salz) trüben die Plausibilität.
5. **ChatGPT 3.5** – Hat die wichtigsten Mengen teils skaliert, aber essenzielle Fehler (viel zu wenig Hefe, kein Öl) und Lücken.
6. **Claude (Anthropic)** – Stimmt in den Mengen vorne und hinten nicht; trotz Extras insgesamt der am wenigsten brauchbare Vorschlag.

Zusammenfassung

Die Auswertung der Pizza-Napoli-Rezepte verschiedener KI-Plattformen (für 503 Portionen, Referenz Betty Bossi) zeigt deutlich, dass **keine KI das Rezept perfekt skaliert und komplett umgesetzt hat**, aber einige doch nahe dran kommen.

Mengenvergleich

Viele Kernzutaten (Mehl, Wasser, Salz im Teig, Tomaten, Anchovies) wurden von den meisten KI-Systemen in der richtigen Grössenordnung angegeben – oft mit moderaten Abweichungen um ±20%. Einzelne Plattformen leisteten sich jedoch extreme Ausreisser: So schlug Perplexity etwa 8-fach zu viel Käse vor, Claude skaliert fast alle Mengen um den Faktor 0,1 (viel zu wenig), und zwei Modelle gaben nur homöopathische Hefemengen an (ein kritischer Fehler für den Teig). Die Standardabweichungen waren bei einigen Zutaten sehr hoch, was auf **inkonsistente Skalierungsfähigkeiten** hindeutet. Insgesamt tendierten die KIs dazu, **Sardellen deutlich geringer** dosieren (wohl realistischer) und **Käse deutlich höher** zu dosieren (mehr nach gängigem Geschmack) als die Referenz.

Vollständigkeit

Alle KI-Rezepte enthielten die Grundelemente (Teigzutaten, Tomatensauce, Käse, Sardellen, Kapern) – in diesem Sinne war jedes Ergebnis brauchbar als Pizza Napoli erkennbar. Allerdings liessen **alle** KIs bestimmte Feinheiten weg, insbesondere **Zwiebeln, Lorbeerblätter und Tomatenpüree** fehlten durchweg. Auch **Zucker** zum Abrunden der Sauce und **Pfeffer** zum Würzen wurden meist ausgelassen. **Oliven** wurden bis auf eine Ausnahme vergessen. Gleichzeitig fügte die Mehrheit der KIs **zusätzliche Kräuter wie Basilikum** hinzu, und die Hälfte würzte mit **Knoblauch**, was im Original nicht der Fall war. Das zeigt, dass die Modelle das Rezept teilweise

„vereinfachten“ (weniger Zutaten) bzw. mit populären Zutaten anreicherten. In Summe war **keine KI 100% vollständig**, aber einige (ChatGPT 4, Perplexity) kamen relativ nah heran, während andere deutliche Lücken aufwiesen.

Qualitative Auffälligkeiten

Die Analyse offenbarte diverse qualitative Probleme:

- **Skalierungsfehler** (z.B. viel zu wenig Hefe bei GPT 3.5/Perplexity, extrem viel Salz bei Grok, falsche Grössenordnungen bei Claude) können Rezepte unbrauchbar machen.
- **Uneinheitliche Rezeptinterpretationen:** Das Spektrum reichte von völliger Würz-Abstinenz (keine Zwiebel, kein Knoblauch, kein Pfeffer bei manchen) bis zu kräftig gewürzten Varianten (mit Knoblauch, reichlich Basilikum, Zucker, etc.).
- **Formulierung und Einheiten:** Einige KI-Angaben waren unpräzise („nach Geschmack“) oder inkonsequent (z.B. Basilikum doppelt genannt), was auf fehlende Verlässlichkeit in der Rezeptausgabe hindeutet.

Bewertung

Am besten schnitt **ChatGPT 4** ab – sein Rezept war bis auf Kleinigkeiten korrekt, umfassend und plausibel. **DeepSeek** lag knapp dahinter, ebenfalls weitgehend vernünftig. **Perplexity** lieferte viele Zutaten und Details, patzte aber bei wichtigen Mengen. **Grok** und **ChatGPT 3.5** zeigten mittlere Leistungen mit einigen gravierenden Auslassungen bzw. Fehlern (besonders GPT 3.5’s Hefemangel). **Claude** landete deutlich auf dem letzten Platz mit inakkuraten Mengen in fast allen Bereichen.

Fazit

Die Untersuchung macht deutlich, dass aktuelle KI-Plattformen zwar in der Lage sind, ein Rezept für eine ungewöhnlich grosse Personenzahl zu generieren, dabei jedoch mit **Inkosistenzen und Auslassungen** zu kämpfen haben. Für eine präzise Skalierung auf z.B. 503 Personen ist weiterhin Vorsicht geboten – die KI-Antworten müssen kritisch überprüft und nachkorrigiert werden. Einige Modelle (allen voran GPT 4) zeigen schon recht zuverlässige Ergebnisse, während andere (Claude) in diesem Testfall versagten. Letztlich kombiniert keine KI perfekt alle Details des Referenzrezepts; die „beste“ Annäherung enthält immer noch Abweichungen, die jedoch – insbesondere bei GPT 4 – relativ gering und behebbar sind.

Sources

Pizza-Rezept-Konsolidiert.xlsx

Prompt für den Report

Analysiere beiliegende Tabelle und erstelle mir eine ausführliche Analysereport. Generieren am Schluss die wichtigsten Erkenntnisse.

Bei der Tabelle handelt es sich dabei um die Ergebnisse, für die Mengen und Zutaten für Pizza Napoli, von verschiedenen KI-Plattformen. Grundrzept stammt von Betty Bossi Webseite für Pizza Napoli für 1 Person. Für den Vergleich mit den Ergebnissen von den verschiedenen KI-Plattformen wurde das Betty Bossi Rezept auf 503 Personen linear hochgerechnet.

Der Prompt für die KI-Plattformen dafür war: "Bitte erstelle ein vollständiges Rezept für Pizza Napoli für 503 Personen." Alles haben den gleichen Prompt erhalten.

Beachte, dass die Einheiten unterschiedlich sind. In der Spalte 'Zusatzinfo' findest du die Umrechnung von Einheiten, wie Prise, Stk., TL oder EL. Um einen Vergleich zu machen, müssen die Einheiten vorerst vereinheitlicht werden.

Brauchst du weitere Informationen?

Napoli für 503 Personen." Alles haben den gleichen Prompt erhalten.

Beachte, dass die Einheiten unterschiedlich sind. In der Spalte 'Zusatzinfo' findest du die Umrechnung von Einheiten, wie Prise, Stk., TL oder EL. Um einen Vergleich zu machen, müssen die Einheiten vorerst vereinheitlicht werden.

Brauchst du weitere Informationen?

Danke für den Kontext und die Datei. Könntest du bitte noch präzisieren, welche Aspekte du im Analysebericht besonders hervorgehoben haben möchtest? Beispielsweise:

1. Soll die Analyse alle Zutaten mengenmässig vergleichen (z.B. Mittelwert, Abweichungen pro Zutat)?
2. Sollen die Abweichungen zu Betty Bossi (Referenz) speziell bewertet werden?
3. Interessieren dich qualitative Auffälligkeiten (z.B. neue Zutaten, fehlende Zutaten)?
4. Möchtest du eine Bewertung der KI-Modelle auf Basis ihrer Resultate?

Mit deinen Antworten kann ich den Analysebericht optimal strukturieren.

1. Soll die Analyse alle Zutaten mengenmässig vergleichen (z.B. Mittelwert, Abweichungen pro Zutat)? Mengenmässig und Vollständigkeit der Zutaten
2. Sollen die Abweichungen zu Betty Bossi (Referenz) speziell bewertet werden? Verwende Betty Bossi Rezept für 503 Personen als Referenz. Das Rezept von Betty Bossi beinhaltet alle Zutaten, die dafür notwendig sind und auch Mengenmässig
3. Interessieren dich qualitative Auffälligkeiten (z.B. neue Zutaten, fehlende Zutaten)? ja
4. Möchtest du eine Bewertung der KI-Modelle auf Basis ihrer Resultate? ja