

Wie genau können LLMs einzelne Befragte in der Marktforschung nachahmen?

Leonard Kinzinger

with Jochen Hartmann (TUM)

Professorship of Digital Marketing
TUM School of Management
Technical University of Munich



Copyright-Hinweis

Bitte beachten Sie, dass die Vortragspräsentation nur für Ihren persönlichen Gebrauch bestimmt ist und nicht vervielfältigt werden darf.

Synthetische Daten bieten neue Lösungen für klassische Marktforschungsprobleme durch Effizienz, Skalierbarkeit und Zugang zu Zielgruppen.



Neue Lösungen

Geschwindigkeit & Kostensparnis

Volumen & Skalierung

Lückenschluss

Erreichbarkeit

Angepasst von: Bannos-Ryback, S., Wang, X (2025). Synthetic Data in Action: Ipsos' Human-Centered Approach to AI Research, MSI/ARF 2025 Analysis & Forecasting

Synthetische Daten bieten ein breites Anwendungsspektrum, das von quantitativer Datenanreicherung bis hin zur interaktiven Simulation von Teilnehmern reicht.

Quantitative Datenanreicherung

(Fokus auf die Verbesserung der Datenbasis)



Data Boosting

Erweiterung bestehender Datensätze durch synthetische Daten, um eine höhere Anzahl simulierter Befragter zu generieren.



Data Imputation

Nutzung synthetischer Daten, um **fehlende Informationen** in unvollständigen Datensätzen statistisch valide **aufzufüllen**.



Data Fusion

Einsatz von Machine Learning zur **Kombination isolierter Datensätze**, um einen neuen, umfassenden Gesamtdatensatz zu erstellen.

Forschung mit synthetischen Teilnehmern

(Fokus auf Simulation)



PersonaBots

Maßgeschneiderte digitale Assistenten, die das **Verhalten** und die Sprache **spezifischer Konsumentensegmente** nachahmen.

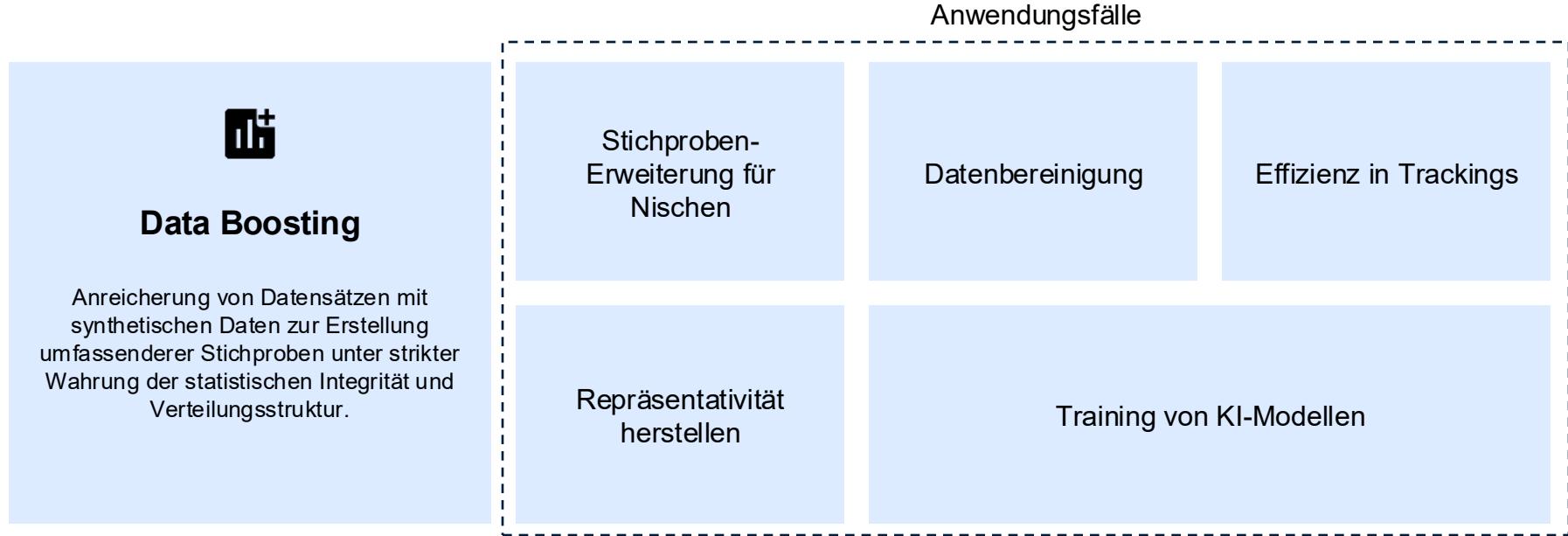


Digital Twins

Erstellung **synthetischer Versionen realer Befragter**, die in der Lage sind, konsistent Antworten auf völlig neue Fragestellungen zu geben.

Angepasst von: Bannos-Ryback, S., Wang, X (2025). Synthetic Data in Action: Ipsos' Human-Centered Approach to AI Research, MSI/ARF 2025 Analysis & Forecasting

Data Boosting erweitert bestehende Datensätze statistisch valide, um Stichproben zu vergrößern, Rauschen zu reduzieren oder Modelle besser zu trainieren.



Angepasst von: Bannos-Ryback, S., Wang, X (2025). Synthetic Data in Action: Ipsos' Human-Centered Approach to AI Research, MSI/ARF 2025 Analysis & Forecasting

Data Imputation nutzt synthetische Daten, um fehlende Informationen intelligent aufzufüllen, die Befragungslast zu senken und Szenarien zu simulieren.



Angepasst von: Bannos-Ryback, S., Wang, X (2025). Synthetic Data in Action: Ipsos' Human-Centered Approach to AI Research, MSI/ARF 2025 Analysis & Forecasting

Data Imputation nutzt synthetische Daten, um fehlende Informationen intelligent aufzufüllen, die Befragungslast zu senken und Szenarien zu simulieren.



PersonaBots

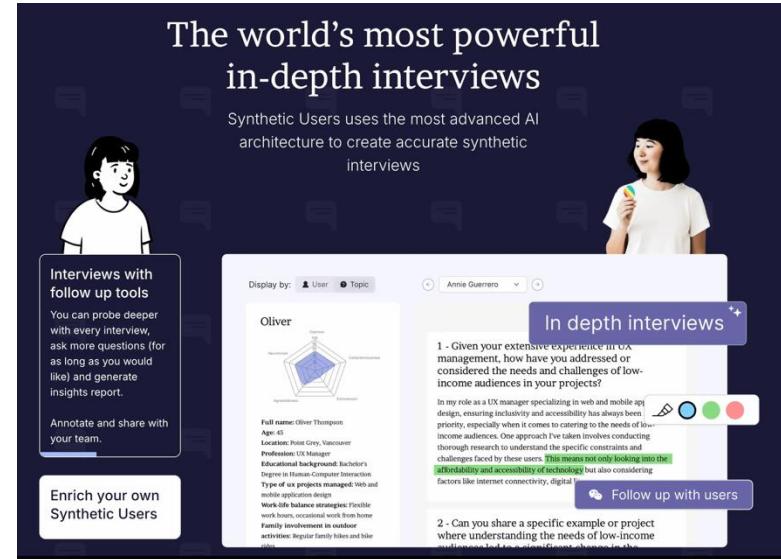
PersonaBots verwandeln statische Zielgruppen-Segmente in interaktive KI-Gesprächspartner, die rund um die Uhr als Sparringspartner verfügbar sind.

- KI-basierte Interaktion
- Authentische Reproduktion
- Strategischer Sparring Partner
- Sofortige Verfügbarkeit

Enrich your own Synthetic Users

The world's most powerful in-depth interviews

Synthetic Users uses the most advanced AI architecture to create accurate synthetic interviews



In depth interviews

1 - Given your extensive experience in UX management, how have you addressed or considered the needs and challenges of low-income audiences in your projects?

2 - Can you share a specific example or project where understanding the needs of low-income audiences has driven your decisions in the

Follow up with users

Angepasst von: Bannos-Ryback, S., Wang, X (2025). Synthetic Data in Action: Ipsos' Human-Centered Approach to AI Research, MSI/ARF 2025 Analysis & Forecasting



Digital Twins sind computergenerierte **Simulationen** realer **Personen**, die mit Hilfe von **LLMs** erstellt werden. Sie basieren auf **echten soziodemografischen** und **psychografischen Daten** und beantworten Fragen so, wie es die repräsentierte Person wahrscheinlich tun würde.

Während PersonaBots aggregierte Eigenschaften ganzer Segmente bündeln, simulieren Digital Twins das Verhalten spezifischer Individuen auf granularer Ebene.



PersonaBots

- **Datenbasis:** Aggregierte Daten von vielen Individuen aus verschiedenen Quellen innerhalb einer Zielgruppe.
- **Funktionsweise:** Bündelung gemeinsamer Merkmale, Verhaltensweisen und Attribute (z.B. Bedürfnisse, Konsumrituale) zu einem repräsentativen Archetyp.
- **Einsatzgebiet:** Ermöglicht naturalistische Interaktionen und qualitative Exploration für Customer Experience und Produktentwicklung.

VS.



Digital Twins

- **Datenbasis:** Tiefe, granulare Daten eines einzelnen Individuums.
- **Funktionsweise:** Präzise Simulation einer spezifischen Entität, um deren einzigartige Reaktionen und Verhaltensweisen in variablen Kontexten vorherzusagen.
- **Einsatzgebiet:** Fokus auf quantitative Tests, Vorhersagen und Personalisierung durch das Erfassen feingliedriger Dynamiken eines einzigen Referenzpunkts.

Klassische Marktforschung ist für viele Anwendungen zu teuer und zu langsam. Könnten synthetische Panels hier unterstützen?

Warum Marktforschung so wichtig ist



Versteht **Lebenswelten** und Wertvorstellungen



Ermöglicht zielgerichtete **Kommunikation**



Reduziert **Missverständnisse** und Streuverluste

Warum sie oft noch nicht eingesetzt wird



Hohe **Kosten** für Studien und Panels



Lange **Durchlaufzeiten**



Zielgruppen schwer erreichbar oder zu klein



Wir untersuchen, ob Large Language Models (**LLM's**) glaubwürdige **Digital Twins** erzeugen können, als **schnelle, günstige Ergänzung** zur klassischen **Marktforschung**.

Einordnung unserer Forschung in die sozialwissenschaftliche und Marketing-Literatur

- Frühere Arbeiten (z. B. **Argyle et al. 2023, Brand et al. 2023**) zeigten, dass LLMs, die auf grobe Demografien konditioniert sind, Umfrage- und Wirtschaftsverhalten replizieren können.
- **Li et al. (2024, MarkSci)** validierten LLMs als Ersatz für menschliche Bewerter bei Aufgaben der Wahrnehmungsanalyse.
- **Park et al. (2024)** untersuchten generative Agenten basierend auf Voice-to-Voice-Interviews.
- **Binz et al. (2025, Nature)** demonstrierten, dass LLMs menschliche Kognition auf Versuchsebene (trial level) vorhersagen können.
- **Toubia et al. (2025, MarkSci)** veröffentlichten Twin-2K-500, einen umfassenden Mikrodatensatz mit über 500 Fragen.

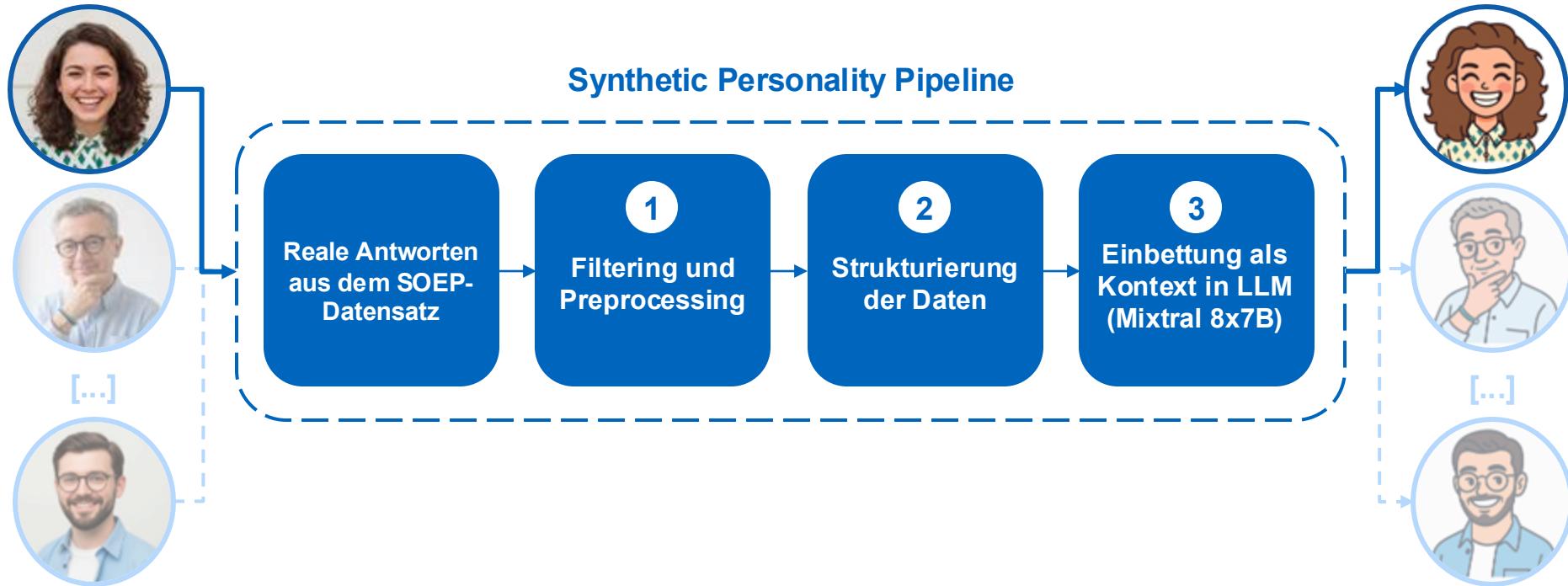


Unsere Forschung erweitert diesen Forschungsstrang, indem wir feingranulare Mikrodaten (SOEP, >800 Fragen) direkt in LLMs einbetten.

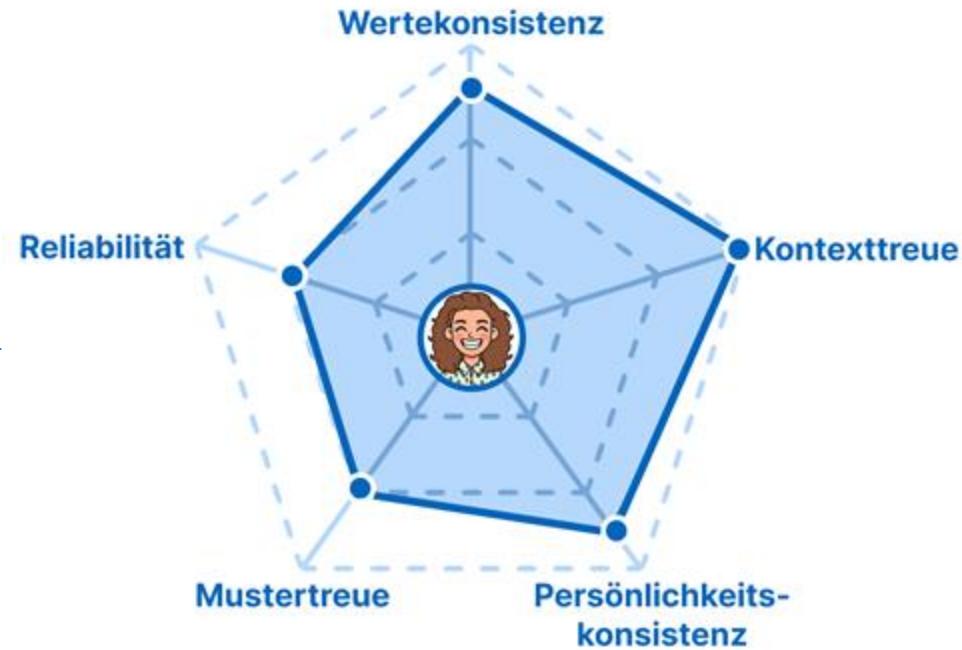
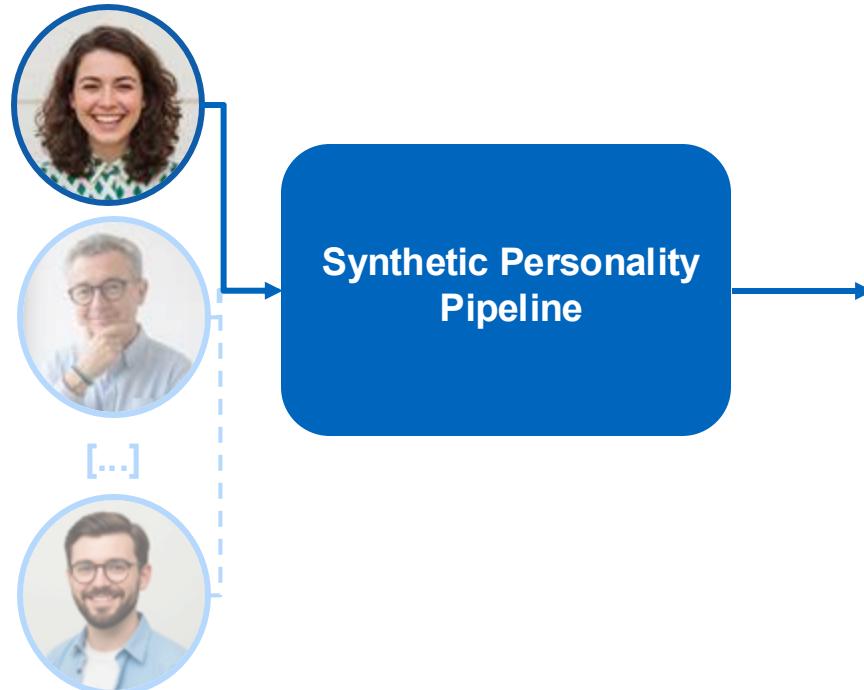
- Geht über Demografie hinaus, um **realistische Personen auf individueller Ebene** zu erfassen.
- Im Gegensatz zu früheren Arbeiten, **keine abstrakten Personas** oder Segmente → stattdessen Simulation von Individuen, die später zu beliebigen Segmenten aggregiert werden können.
- Direkt relevant für **Marktforschungsinstitute** (die solche Mikrodaten bereits besitzen) → ermöglicht neue Umfragesimulationen ohne kostspielige Feldarbeit.

Teil 1: Wie erstellt man Digital Twins?

Wir erzeugen **synthetische Persönlichkeiten**, indem wir **reale Daten** dieser Person strukturieren und gezielt in **LLMs** einbetten.



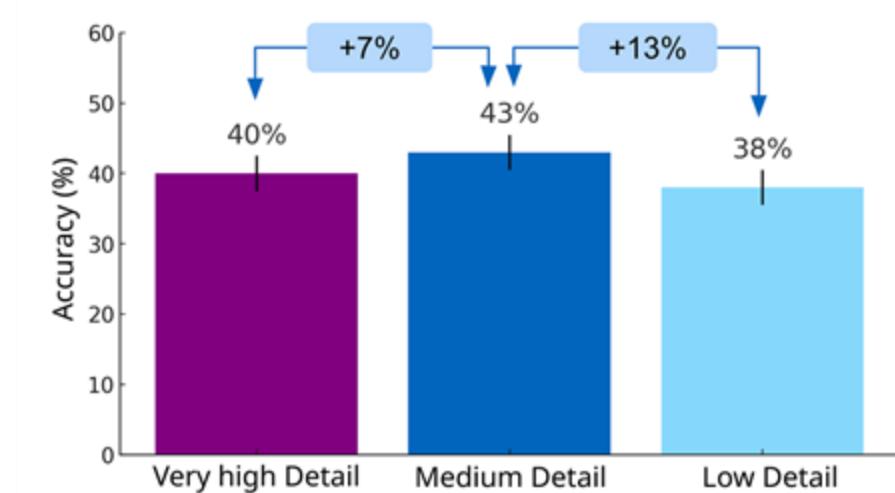
Wir bewerten die Qualität dieser synthetischen Persönlichkeiten anhand von fünf zentralen Dimensionen.



Die Evaluationsdimensionen bauen auf früheren Arbeiten auf: Inspiriert von Argyle et al. (2023) zur algorithmischen Treue (algorithmic fidelity), erweitert um Kriterien aus der Rollenspielforschung (Shao et al. (2023), Wang et al. (2023)).

Experiment 1 : Eine **mittlere Kontexttiefe** liefert die qualitativ besten Antworten.

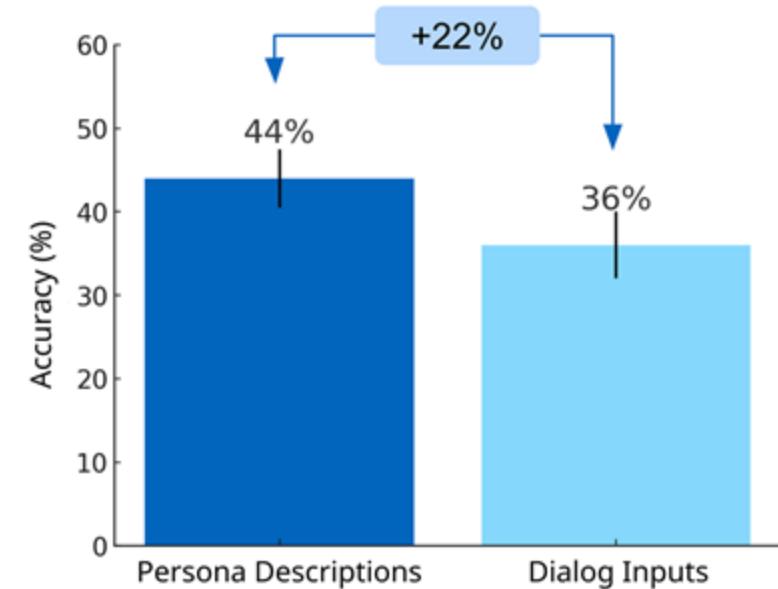
- Bestes Ergebnis mit **mittlerem Detaillierungsgrad** (~400 Datenpunkte).
- **Zu wenig**: Führt zu vagen, generischen Antworten.
- **Zu viel**: Führt zu **Ablenkung** (Distraction) & wichtige Informationen gehen verloren.
- **Annahme**: Stärkere Effekte bei noch neueren LLMs zu erwarten (z. B. GPT-5).



Hinweis: Die Leistung wurde an einem Hold-out-Set evaluiert, das von der Erstellung der synthetischen Persönlichkeit ausgeschlossen war.

Experiment 2 : Personenbeschreibungen führen zu präziseren Antworten als strukturierte Dialog-Inputs.

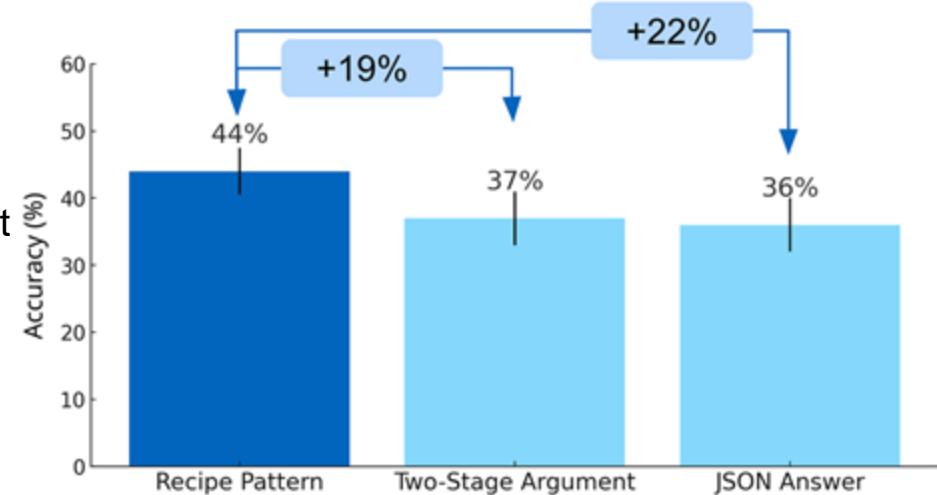
- **Personenbeschreibungen** liefern eine höhere Genauigkeit als Frage-Antwort-Dialoge (Q&A) als Input.
- **LLMs** verstehen Inputs mit **höherer Informationsdichte** besser.
- **Annahme:** Input in natürlicher Sprache liefert wertvolle, **zusätzliche Informationen** (impliziter Kontext).



Hinweis: Die Leistung wurde an einem Hold-out-Set evaluiert, das von der Erstellung der synthetischen Persönlichkeit ausgeschlossen war.

Experiment 3 : Rezept-prompts aktivieren den Denkprozess (Test-Time-Reasoning) und verbessern die Antwortqualität.

- **Mehrere Aufgaben gleichzeitig** (z. B. Antworten + Formatieren) verschlechtern die Qualität.
- Mehr **Bedenkzeit** (Thinking Time / TTC) führt zu besseren, reflektierteren Antworten.
- Offene **Rezept-Prompts** übertreffen enge Pro/Contra-Strukturen.



Hinweis: Die Leistung wurde an einem Hold-out-Set evaluiert, das von der Erstellung der synthetischen Persönlichkeit ausgeschlossen war.

Teil 2: Wie genau sind die Antworten von Digital Twins?

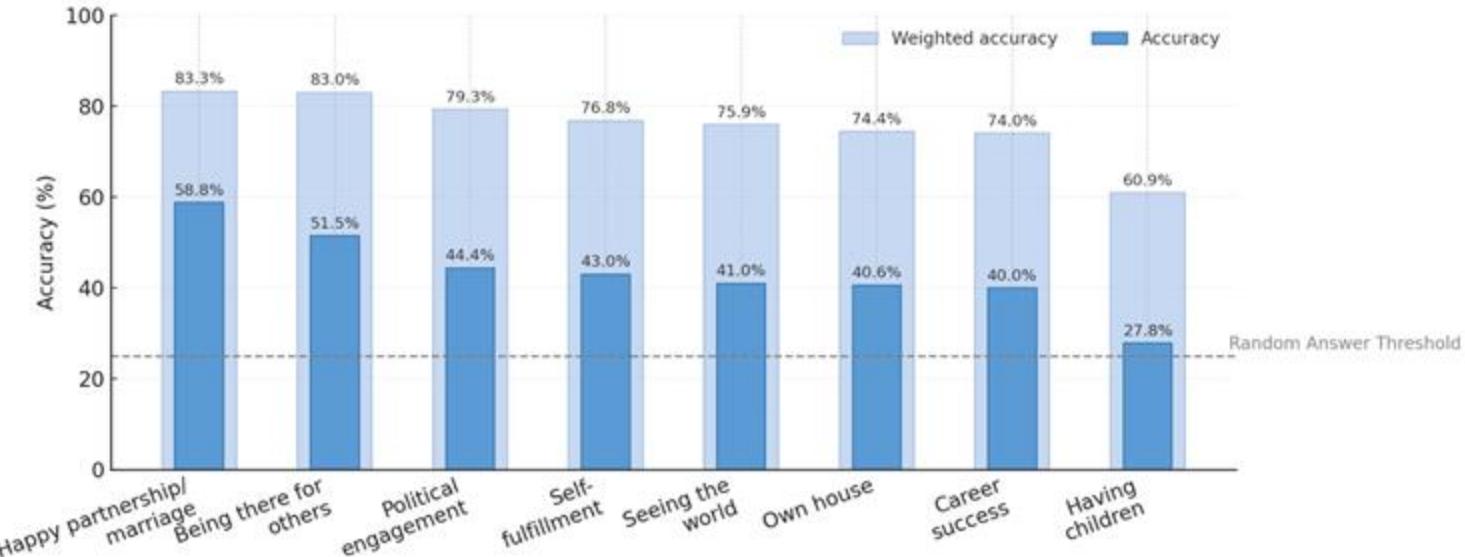
Digital Twins zeigen bereits in vielen Dimensionen eine hohe Leistung, offenbaren jedoch noch spezifische Optimierungspotenziale.

-  **Erinnerungsvermögen:** 92–98% Genauigkeit → zuverlässiger Abruf von Fakten, die in ihrer Beschreibung eingebettet sind.
-  **Wertekonsistenz:** ~43% direkte, 77% gewichtete Übereinstimmung → Antworten liegen meist innerhalb einer Stufe der wahren Antwort, aber noch Verbesserungspotenzial.
-  **Reliabilität:** SD 0,3–0,6 → Antworten schwanken typischerweise nur zwischen zwei benachbarten Kategorien, was dem menschlichen Antwortverhalten entspricht.
-  **Persönlichkeitskonsistenz:** Big Five-Merkmale werden gut reflektiert → Nuancen zwischen Individuen werden akkurat erfasst (insbesondere bei Extraversion & Neurotizismus).
-  **Mustertreue:** Es verbleiben Verzerrungen → progressive Milieus sind überrepräsentiert, traditionelle Milieus unterrepräsentiert.

Wertekonsistenz: Synthetische Persönlichkeiten bilden menschliche Wertemuster bereits erstaunlich gut ab.

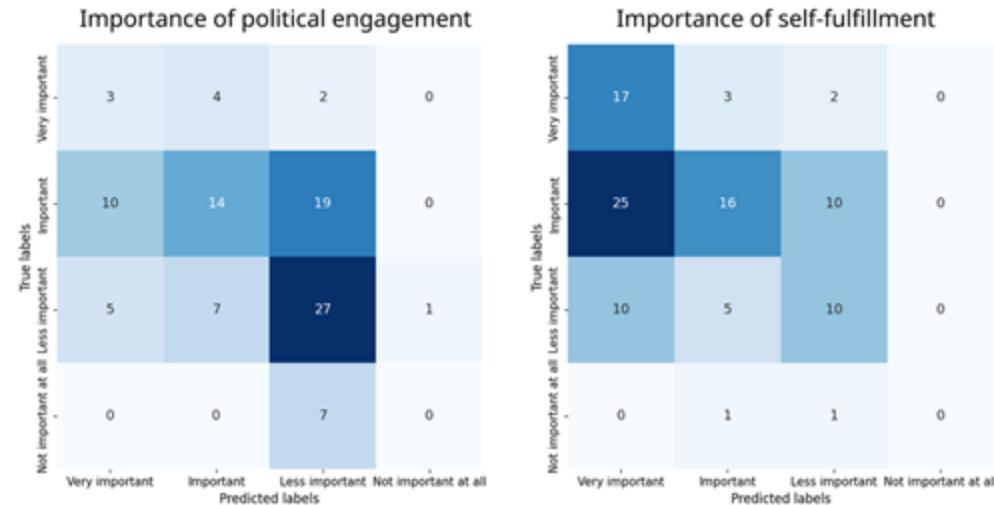
- **Durchschnittliche Genauigkeit** (Accuracy) über 8 Wertefragen hinweg: **43,4 %**.
- Durchschnittliche gewichtete Genauigkeit (Weighted Accuracy): 77 %.

→ **Erkenntnis:** Antworten der synthetischen Persönlichkeiten liegen meist innerhalb einer Stufe der wahren Werte (z. B. "Wichtig" statt "Sehr wichtig").



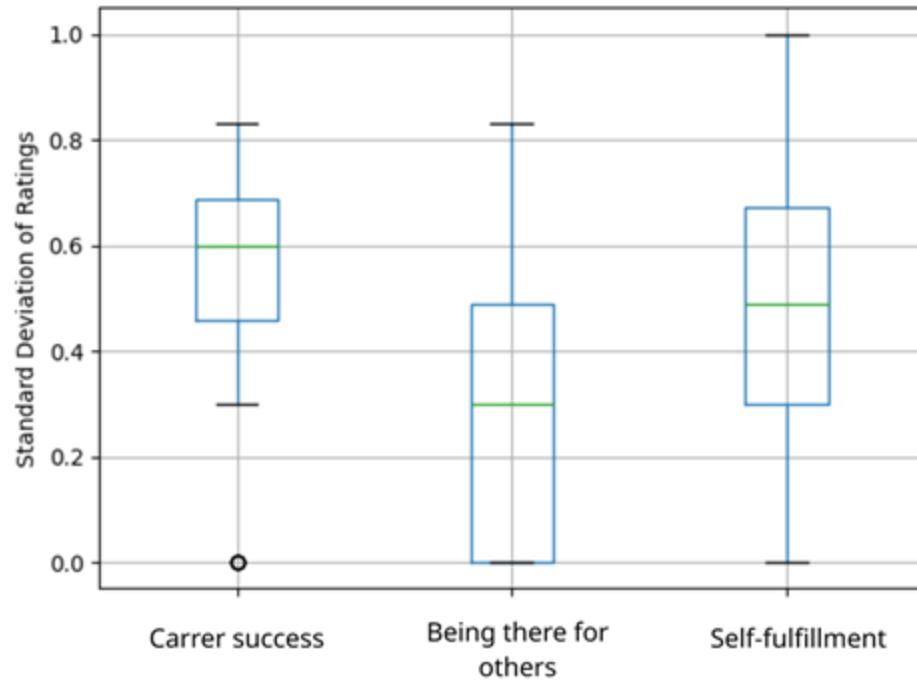
Detailanalyse Werte: Während die Verteilung meist gut getroffen wird, zeigen sich bei spezifischen Themen systematische Verzerrungen.

- Bei den meisten Evaluationsfragen erfassen synthetische Persönlichkeiten die **gesamte Verteilung** der Antworten über alle Wichtigkeitsstufen hinweg.
- Die Wichtigkeit, **Kinder zu haben**, wird nicht korrekt erfasst (**27,8 % Genauigkeit**) – die Modelle gehen bei älteren Befragten fälschlicherweise von Irrelevanz aus, während echte Menschen hier eine **breitere Lebensperspektive** einnehmen.



Reliabilität: Die Antwortvarianz synthetischer Persönlichkeiten ähnelt der natürlicher menschlicher Schwankungen.

- Getestet mit **Mixtral 8x7B** bei Standard-Temperatur (0,7).
- Antworten wurden **10x pro Frage wiederholt**, um die Varianz zu messen.
- Die mittleren **Standardabweichungen** liegen zwischen **0,3 – 0,6** (was typischer menschlicher Inkonsistenz bei wiederholten Befragungen entspricht).



Reliabilität: Detaillierte Analyse der Antwortvarianz zeigt menschliche Schwankungsmuster.

- Getestet mit **Mixtral 8x7B** bei Standard-Temperatur (0,7).
- Antworten wurden **10x pro Frage wiederholt**, um die Varianz zu messen.
- Die mittleren **Standardabweichungen** liegen zwischen 0,3 – 0,6.

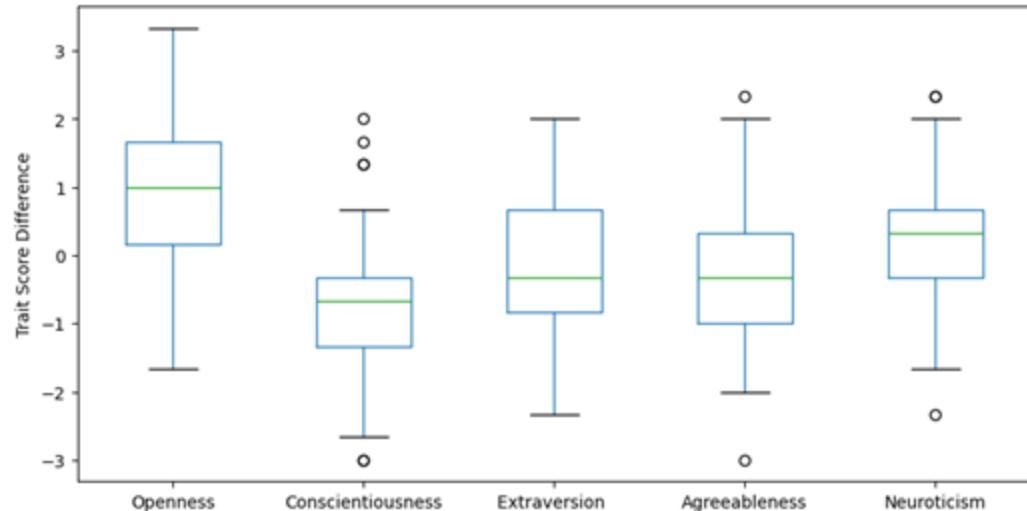
0	1	2	3	4	5	6	7	8	9	Standard Deviation
Responses with Highest Standard Deviation										
1	3	3	2	2	1	0	3	3	2	1.0
3	3	1	1	1	3	1	3	3	1	1.0
1	1	1	3	3	3	3	3	2	1	0.9
Responses with Median Standard Deviation										
3	2	2	2	2	2	3	3	2	3	0.5
3	3	2	2	3	2	3	2	3	3	0.5
2	1	1	1	2	2	2	1	1	1	0.5
Responses with Lowest Standard Deviation										
3	3	3	3	3	3	3	3	3	3	0.0
3	3	3	3	3	3	3	3	3	3	0.0
3	3	3	3	3	3	3	3	3	3	0.0



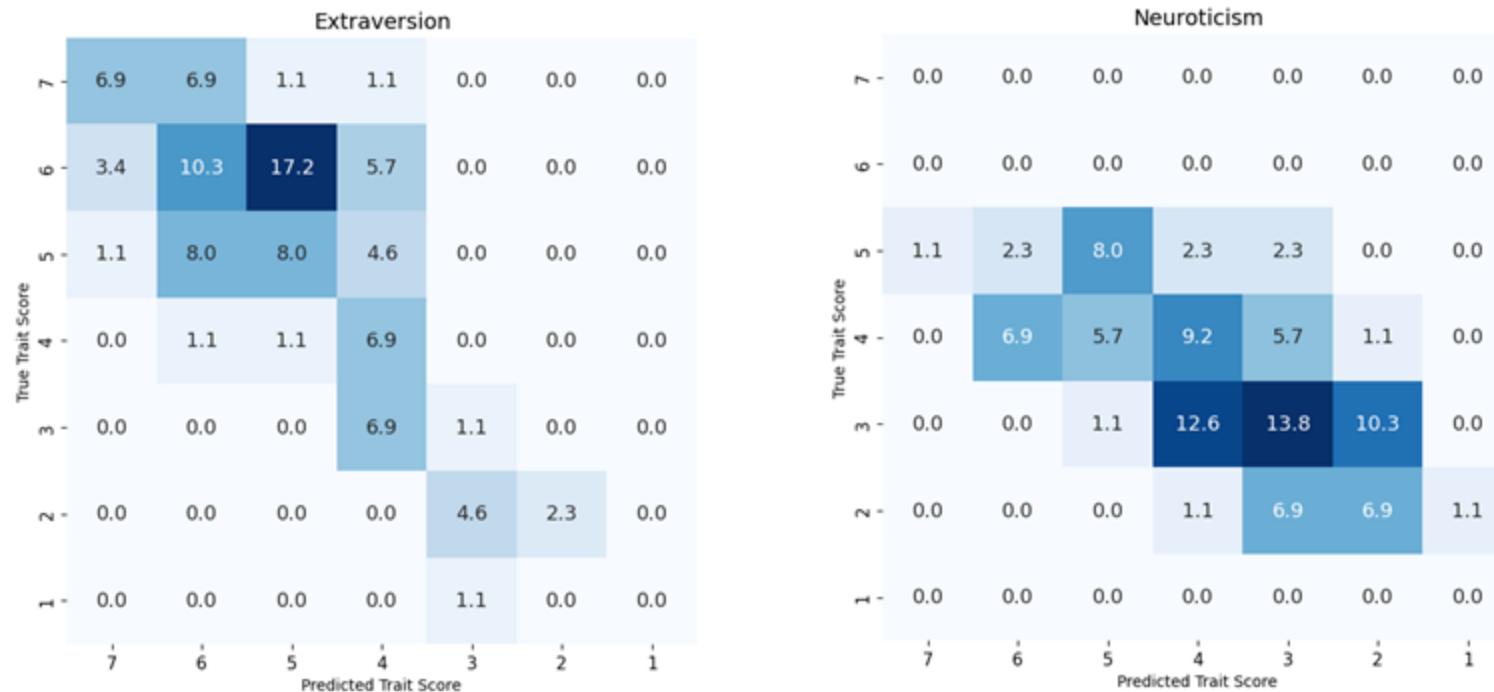
Die Variabilität entspricht menschlichen Antwortmustern (keine statische Wiederholung, aber Konsistenz im Kern).

Konsistenz der Persönlichkeit: Synthetische Persönlichkeiten spiegeln Persönlichkeitsmerkmale weitgehend genau wider.

- BFI-S-Test mit Big Five Traits (SOEP vs. Synth) zeigt **gute Übereinstimmung**
- Besonders **Neurotizismus** und **Extraversion** werden treffsicher vorhergesagt (RMSE < 1)
- **Openness** wird systematisch **überschätzt**, **Conscientiousness** **unterschätzt**



Besonders bei den Merkmalen **Extraversion** und **Neuroticismus** gibt es eine hohe Übereinstimmung.



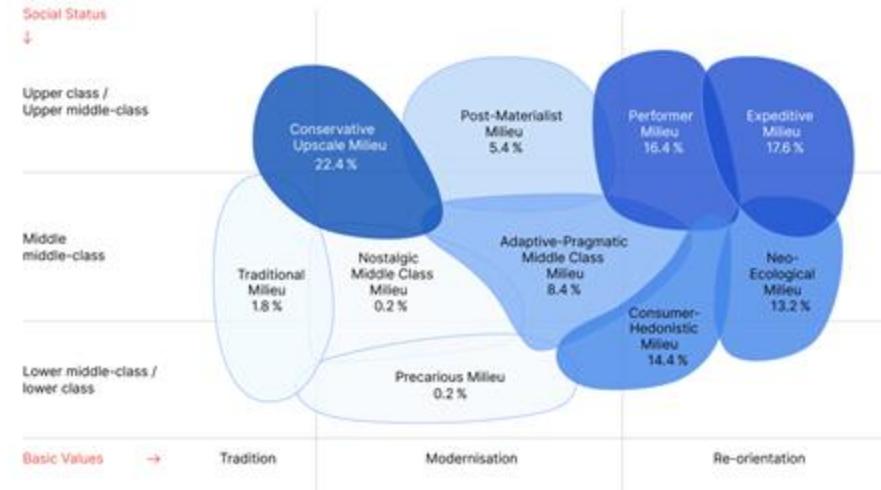
Synthetische Persönlichkeiten spiegeln **progressive Werte** **überproportional** ab und verzerrten die Milieuverteilung.

Distribution of Sinus-Milieus in Germany



Sinus-Milieus® in Germany: The potato diagram shows the distribution of the Sinus-Milieus® in the German population. Source: Sinus Institute (2024). The gold standard for target group segmentation. Retrieved April 25, 2024, from <https://www.sinus-institut.de/en/sinus-milieus>

Distribution of Sinus-Milieus in Synthetic Panel



Sinus-Milieus® in our Synthetic Panel. Adjusted with different values from: Sinus Institute (2024). The gold standard for target group segmentation. Retrieved April 25, 2024, from <https://www.sinus-institut.de/en/sinus-milieus>

Zusammenfassung der zentralen Beiträge, Ergebnisse und Optimierungspotenziale

Beiträge & Ergebnisse

- Erstmalige Generierung von Digital Twins auf Basis realer Mikrodaten (SOEP)
- Framework zur **systematischen Evaluation** über fünf Qualitätsdimensionen hinweg.
- Hohe Konsistenz in **Persönlichkeit**, **Erinnerungsvermögen** und **Antwort-Reliabilität**.
- Verzerrung (Bias) hin zu **progressiven Werten**

Optimierungspotenziale

- Einsatz **leistungsfähigerer Modelle** mit besserem **Kontextverständnis**
- Verwendung von **Input in natürlicher Sprache** anstelle strukturierter Frage-Antwort-Daten
- **Pre-Training** oder **Instruction-Tuning** auf **marktforschungsspezifischen Datensätzen**

Potenzielle Anwendungsgebiete

- **Pre-Testing von Umfragen** ohne kostspielige Feldarbeit.
- Testen von **Werbung** und **Produktkonzepten** an synthetischen Panels.
- **Szenario-Analyse**: Vorhersage, wie verschiedene Bevölkerungsgruppen reagieren könnten.
- Unterstützung von **Längsschnittstudien**, bei denen die erneute Kontaktierung von Teilnehmern schwierig ist.

Vielen Dank für Ihre Aufmerksamkeit!

Leonard Kinzinger

Doktorand @ TUM, Professorship of Digital Marketing

LinkedIn: [@leonardkinzinger](https://www.linkedin.com/in/leonardkinzinger)

Email: leonard.kinzinger@tum.de



Let's connect on
LinkedIn

Backup

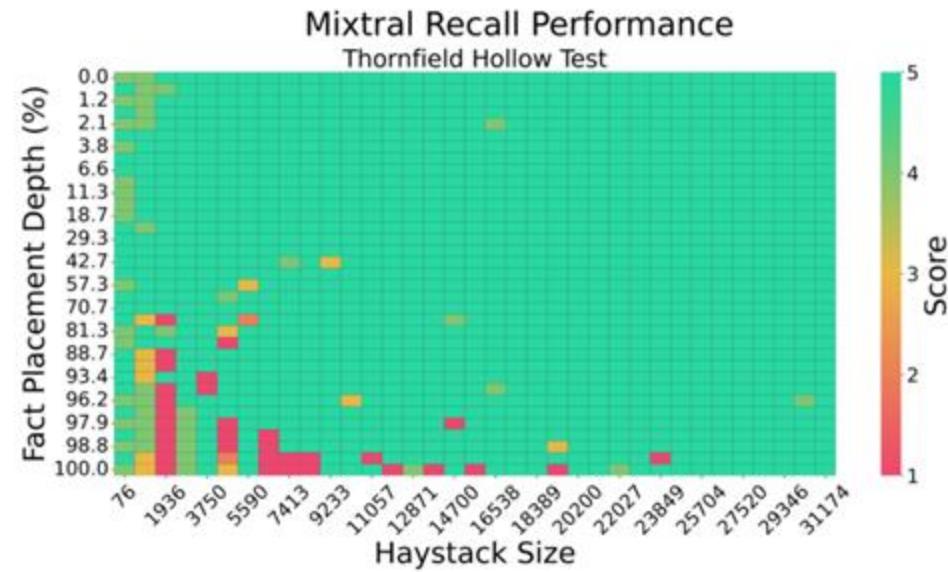
Memorisation: Synthetic personalities **remember embedded information** with high accuracy.

- **92–98% accuracy** in almost all areas of the persona description
- Errors in the middle of the input due to **position effects** & **preprocessing errors**, not due to retrieval
- Retrieval itself worked reliably; with correct input, hardly any dropouts

Categories (Grouped by position within the descriptions)	Number of Questions	Mean Memorisation Accuracy (%)	Standard Deviation
1-4	9	98.4%	0.026
5-8	9	94.9%	0.038
9-12	12	60.6%	0.204
13-16	6	92.5%	0.099
17-20	5	96.9%	0.063

Memorisation: Synthetic personalities **remember embedded information** with high accuracy.

- **92–98% accuracy** in almost all areas of the persona description
- Errors in the middle of the input due to **position effects** & **preprocessing errors**, not due to retrieval
- Retrieval itself worked reliably; with correct input, hardly any dropouts



Needle-in-the-haystack Results for Mixtral 8x7B: Machlab and Battle, 2024 showed that the recall ability of the Mixtral 8x7B model is dependent on the position within and length of the context window. Source: Machlab and Battle, 2024, pp. 5–6