Usage of Large Language Models as **Diagnostic Tools** in Psychiatric Interviews

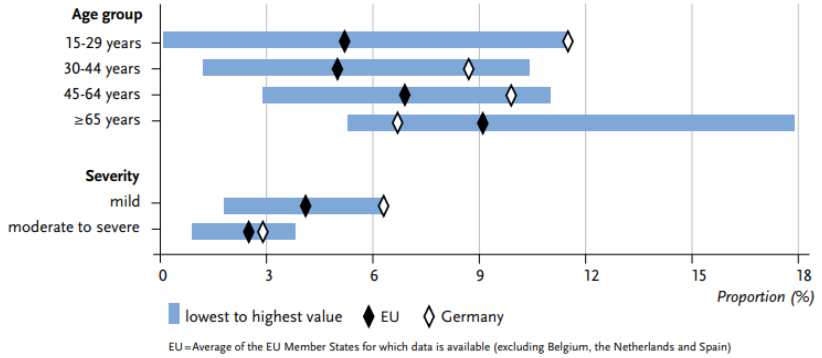
A.C. Joblin, M. Schiltenwolf, D. Fürstenau & S. Schreiter

Retreat Kloster Lehnin 2025

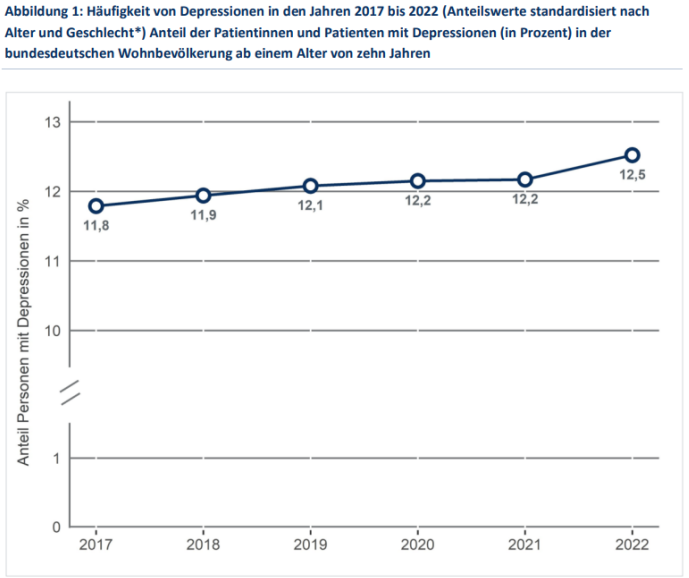
Mental health and economic pressure.

- 12,5 % of the population Germany was diagnosed with a depression within the last 5 years.
- Yearly costs of each depression patient are estimated between 400-3300 €
- There is extensive economic pressure on healthcare providers while the workload on healthcare practitioners is rising.

→ Which tools can support practitioners in their daily work?

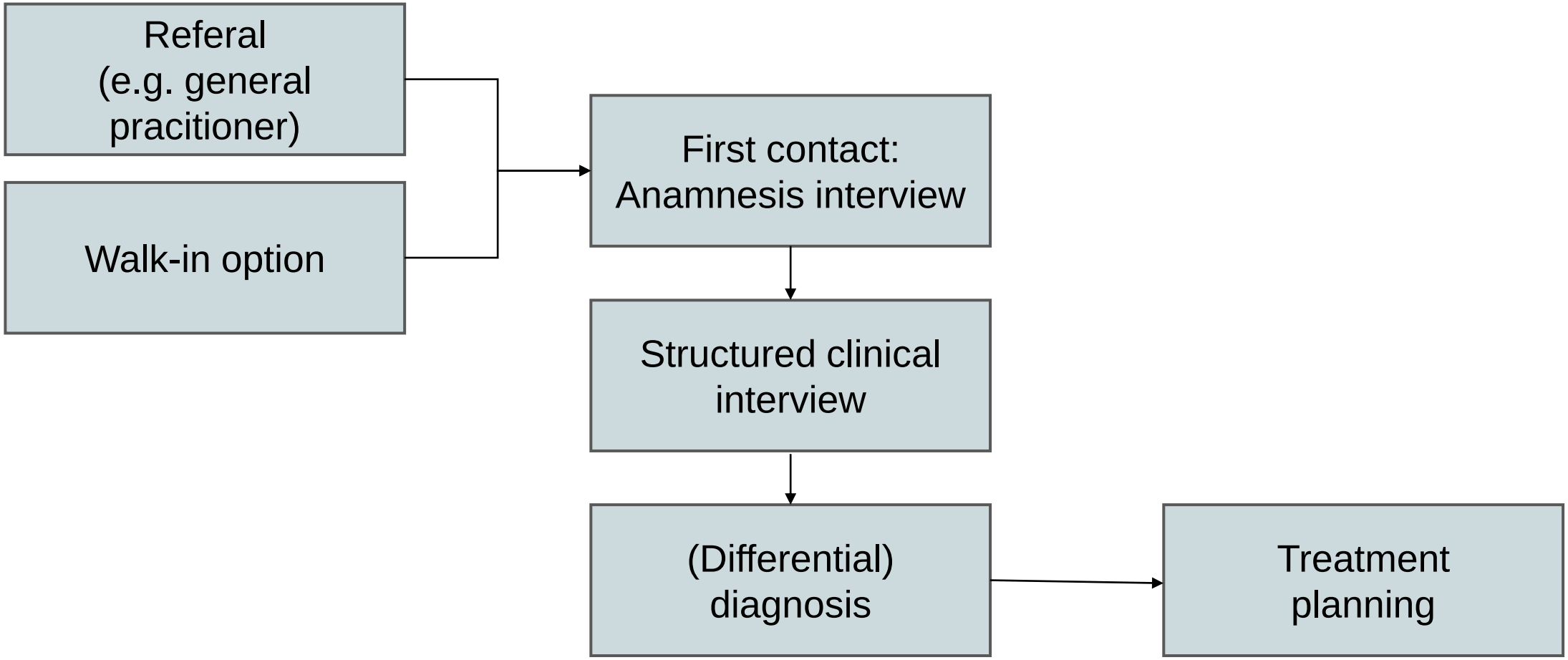


Hapke et al. (2019)

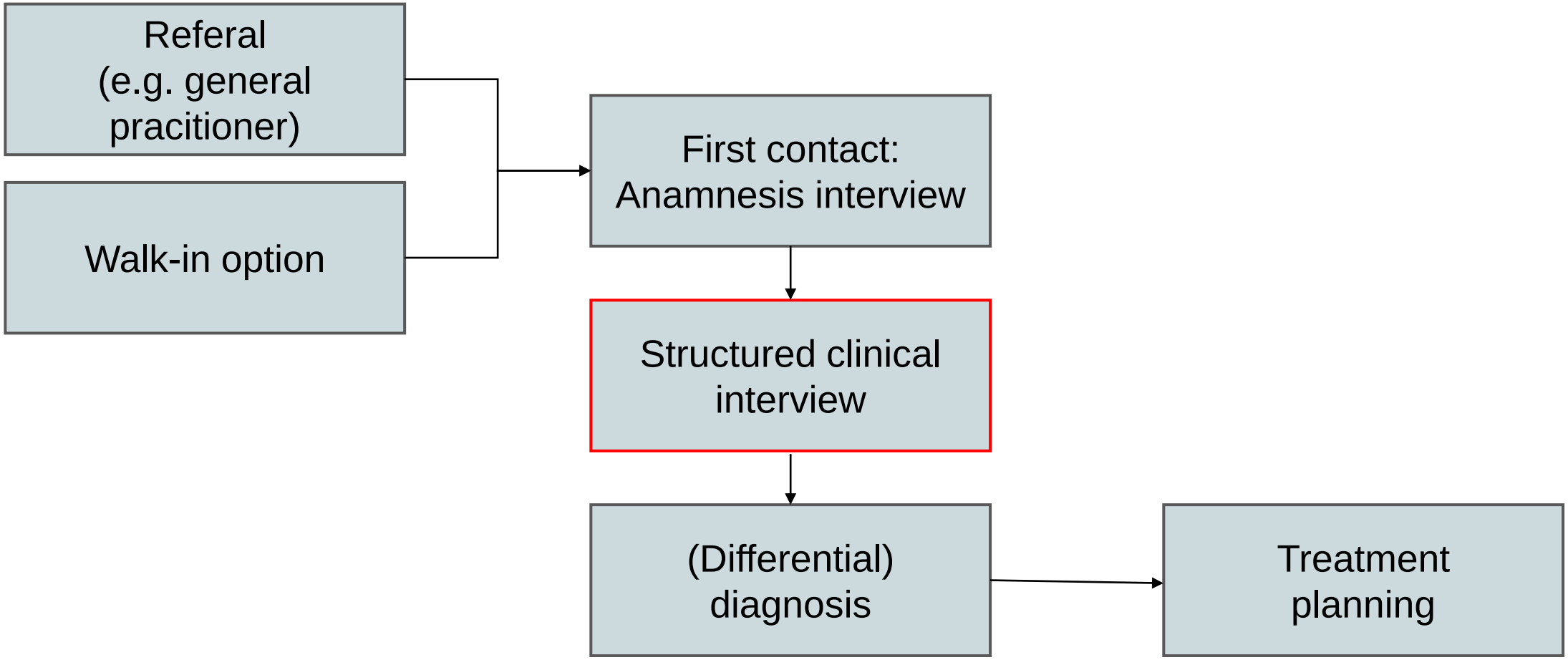


Gesundheitsatlas Deutschland (2024)

First contact patient - mental health practitioner



First contact patient - mental health practitioner



Structured Clinical Interview for DSM-V (SCID-V)

Structured interview for making diagnosis according to the American DSM-V.

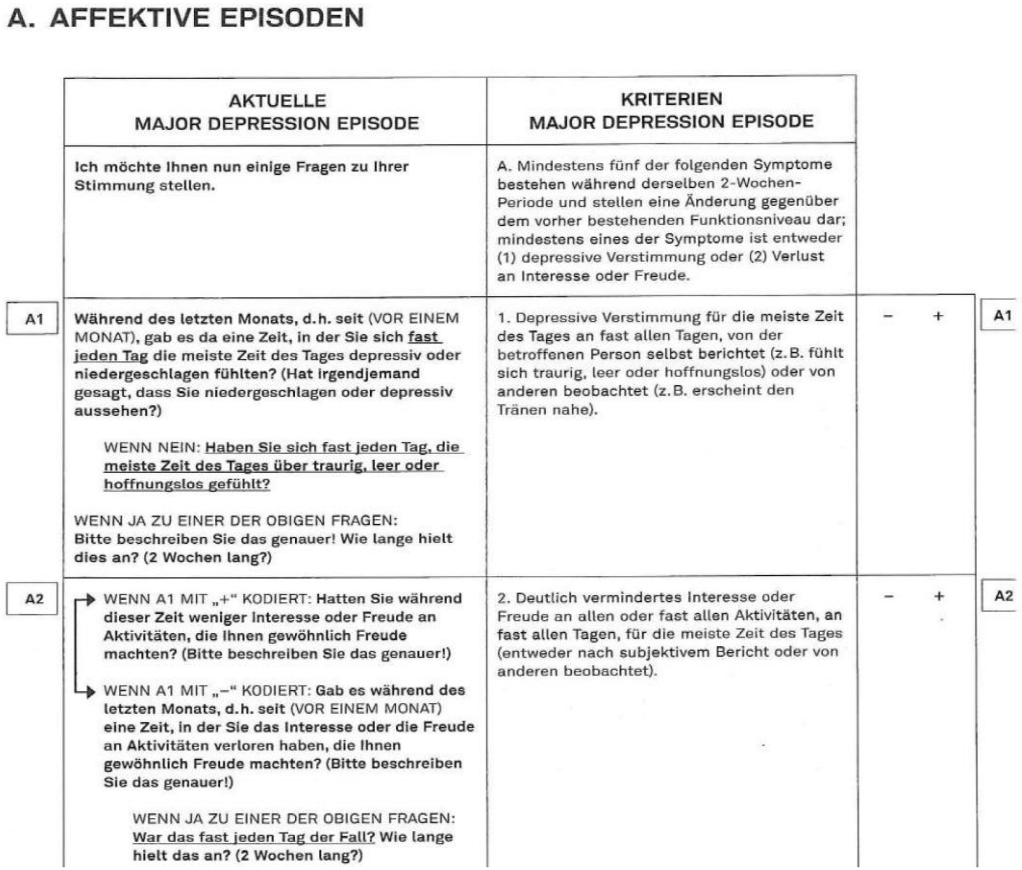
Performed by **diagnostic specialists** (usually psychologists or physicians).

Duration >1h.

Five axes:

1. **Clinical disorders**
2. Personality disorders
3. General medical conditions
4. Psychosocial and environmental problems
5. Global assessment of functioning

Structured Clinical Interview for DSM-V (SCID-V)



A. AFFEKTIVE EPISODEN

	AKTUELLE MAJOR DEPRESSION EPISODE	KRITERIEN MAJOR DEPRESSION EPISODE			
	Ich möchte Ihnen nun einige Fragen zu Ihrer Stimmung stellen.	A. Mindestens fünf der folgenden Symptome bestehen während derselben 2-Wochen-Periode und stellen eine Änderung gegenüber dem vorher bestehenden Funktionsniveau dar; mindestens eines der Symptome ist entweder (1) depressive Verstimmung oder (2) Verlust an Interesse oder Freude.			
A1	**Während des letzten Monats, d.h. seit** (VOR EINEM MONAT), **gab es da eine Zeit, in der Sie sich fast jeden Tag die meiste Zeit des Tages depressiv oder niedergeschlagen fühlten? (Hat irgendjemand gesagt, dass Sie niedergeschlagen oder depressiv aussehen?)** WENN NEIN: **Haben Sie sich fast jeden Tag, die meiste Zeit des Tages über traurig, leer oder hoffnungslos gefühlt?** WENN JA ZU EINER DER OBIGEN FRAGEN: **Bitte beschreiben Sie das genauer! Wie lange hielt dies an? (2 Wochen lang?)**	1. Depressive Verstimmung für die meiste Zeit des Tages an fast allen Tagen, von der betroffenen Person selbst berichtet (z.B. fühlt sich traurig, leer oder hoffnungslos) oder von anderen beobachtet (z.B. erscheint den Tränen nahe).	–	+	A1
A2	WENN A1 MIT „+" KODIERT: **Hatten Sie während dieser Zeit weniger Interesse oder Freude an Aktivitäten, die Ihnen gewöhnlich Freude machten? (Bitte beschreiben Sie das genauer!)** WENN A1 MIT „–" KODIERT: **Gab es während des letzten Monats, d.h. seit** (VOR EINEM MONAT) **eine Zeit, in der Sie das Interesse oder die Freude an Aktivitäten verloren haben, die Ihnen gewöhnlich Freude machten? (Bitte beschreiben Sie das genauer!)** WENN JA ZU EINER DER OBIGEN FRAGEN: **War das fast jeden Tag der Fall? Wie lange hielt das an? (2 Wochen lang?)**	2. Deutlich vermindertes Interesse oder Freude an allen oder fast allen Aktivitäten, an fast allen Tagen, für die meiste Zeit des Tages (entweder nach subjektivem Bericht oder von anderen beobachtet).	–	+	A2

In the present study we focus on axis 1A: ***Affective Episodes***.

Identification of past and/or ongoing:

- Major Depressive Episodes
- Manic Episodes
- Hypomanic Episodes
- Mixed Episodes

Possible diagnosis include:

- Major Depression
- Bipolar Disorder I & II

Challenge differential diagnosis
Major Depression - Bipolar Disorder

Feature	Major Depression	Bipolar I Disorder	Bipolar II Disorder
Manic Episode	No	Yes	No
Hypomanic Episode	No	May occur, but not required	Yes
Major Depressive Episode	Yes	Common, but not required	Yes
Duration of Episode	≥ 2 weeks	≥ 1 week (or any duration if hospitalized)	≥ 4 consecutive days (hypomania) + ≥ 2 weeks (depression)
Functional Impairment	Significant distress or impairment	Severe impairment or hospitalization or psychosis	Impairment mainly due to depression, not hypomania
Psychotic Features	May occur in severe cases	May occur during mania	Not present in hypomania
Exclusion of Other Disorders	No history of mania/hypomania; not better explained by psychosis	Must rule out other causes of mood elevation	No manic episodes; not explained by another condition

→ There is **symptom overlap** between the disorders.
→ Often Bipolar Disorders remain **hidden** and are **misdiagnosed** as Major Depressions.
→ **9 years** is the average time between initial clincial presentation and accurate diagnosis of bipolard disorder.

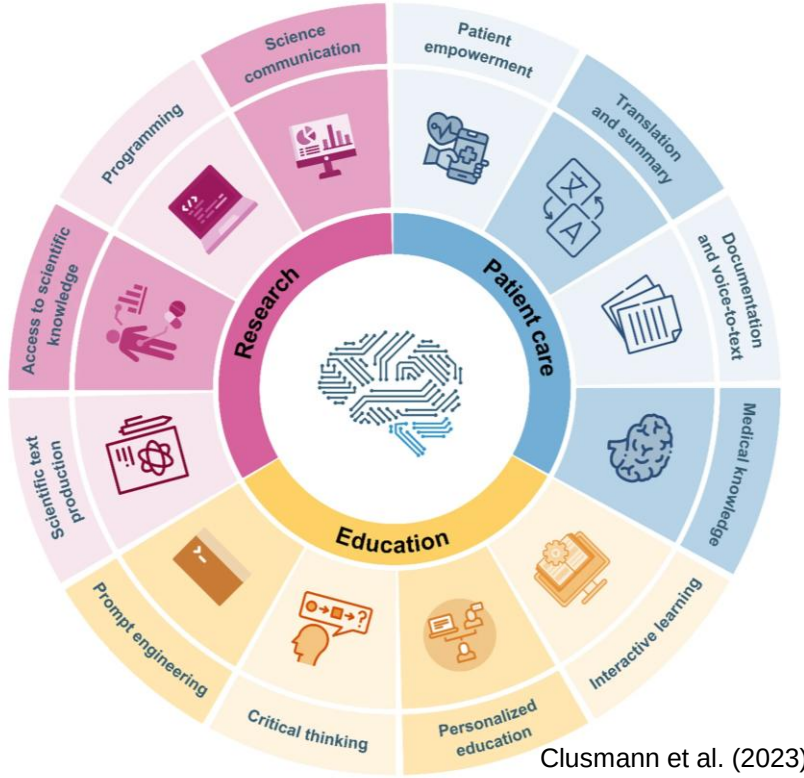
LLMs in (mental) healthcare Promising first results

LLMs are discussed as helpful in

- Clinical documentation
- Workflow optimazation
- Education for personal and patients

But can LLMs serve as decision support systems?

- A range of models were trained on mental health specific data sets (e.g. from mental health specific discussion on Reddit.
- Even using out-of-the box, zero-shot prompted GPT-4 models are quite good in categorising self-descriptions as clinically relevant or not (balanced accuracy scores 0.75).



Clusmann et al. (2023)

Dataset	Task	Dataset Size	Text Length (Token)
Dreaddit [120] *Source: Reddit*	#1: Binary Stress Prediction *post-level*	Train: 2838 (47.6% False, 52.4% True) Test: 715 (48.4% False, 51.6% True)	Train: 114 ± 41 Test: 113 ± 39
DepSeverity [80] *Source: Reddit*	#2: Binary Depression Prediction *post-level*	Train: 2842 (72.9% False, 17.1% True) Test: 711 (72.3% False, 17.7% True)	Train: 114 ± 41 Test: 113 ± 37
	#3: Four-level Depression Prediction *post-level*	Train: 2842 (72.9% Minimum, 8.4% Mild, 11.2% Moderate, 7.4% Severe) Test: 711 (72.3% Minimum, 7.2% Mild, 11.5% Moderate, 10.0% Severe)	Train: 114 ± 41 Test: 113 ± 37
SDCNL [49] *Source: Reddit*	#4: Binary Suicide Ideation Prediction *post-level*	Train: 1516 (48.1% False, 51.9% True) Test: 379 (49.1% False, 50.9% True)	Train: 101 ± 161 Test: 92 ± 119
CSSRS-Suicide [40] *Source: Reddit*	#5: Binary Suicide Risk Prediction *user-level*	Train: 400 (20.8% False, 79.2% True) Test: 100 (25.0% False, 75.0% True)	Train: 1751 ± 2108 Test: 1909 ± 2463
	#6: Five-level Suicide Risk Prediction *user-level*	Train: 400 (20.8% Supportive, 20.8% Indicator, 34.0% Ideation, 14.8% Behavior, 9.8% Attempt) Test: 100 (25.0% Supportive, 16.0% Indicator, 35.0% Ideation, 18.0% Behavior, 6.0% Attempt)	Train: 1751 ± 2108 Test: 1909 ± 2463
Red-Sam [105] *Source: Reddit*	#2: Binary Depression Prediction *post-level*	External Evaluation: 3245 (26.1% False, 73.9% True)	External Evaluation: 151 ± 139
Twt-60Users [56] *Source: Twitter*	#2: Binary Depression Prediction *post-level*	External Evaluation: 8135 (90.7% False, 9.3% True)	External Evaluation: 15 ± 7
SAD [76] *Source: SMS-like*	#1: Binary Stress Prediction *post-level*	External Evaluation: 6185 (6.0% False, 94.0% True)	External Evaluation: 13 ± 6

Xu et al. (preprint)

But can LLMs be utilized as diagnostic tools?

In this study we aimed to to study how LLMs perform if used in the the regular diagnosis process using data from the SCID-V (Axis A).

1. Can an LLM identify diagnosis-relevant text section in a clinical interview?
2. Can an LLM identify the correct symptoms described in such text sections?
3. Is the LLM based section/symptom estimation worse than the estimation of clinical experts?

Study plan

Dataset: 50 anonymized SCID-5 CV patient interviews (Charité-Psychiatry; focus depression/ bipolar)

+ 30 healthy controls

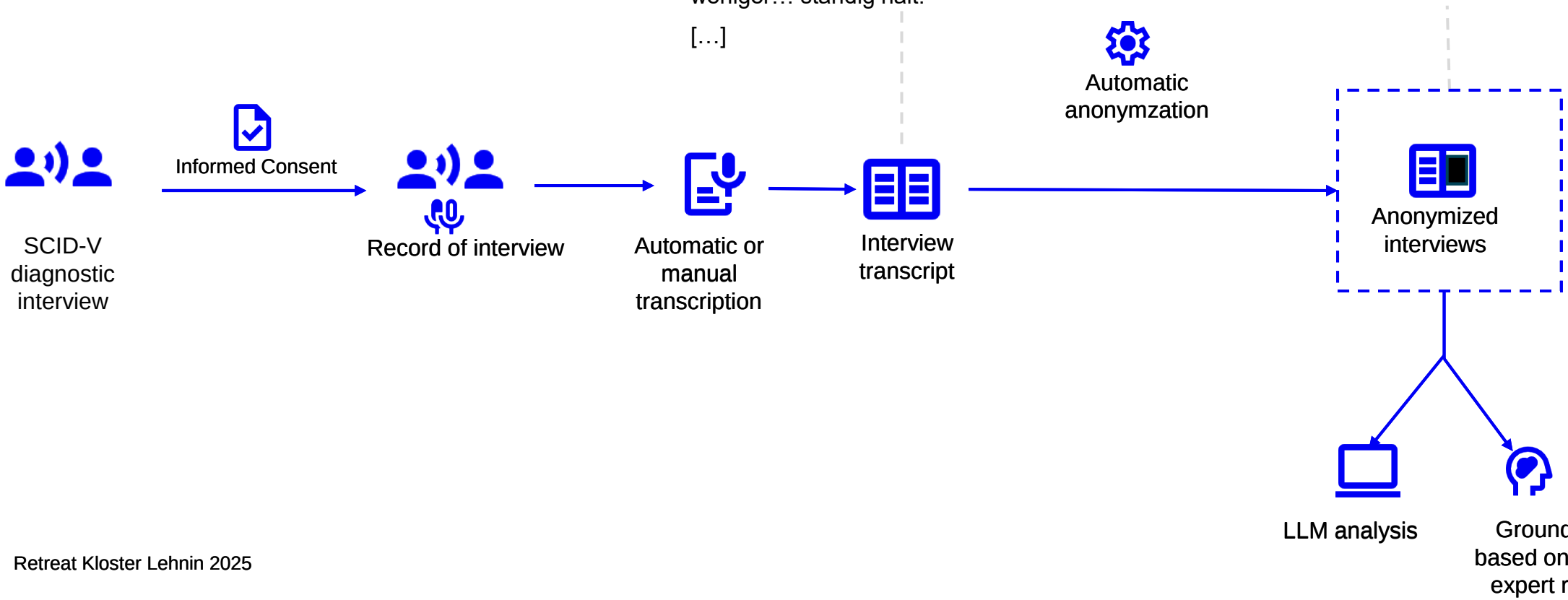
Arzt: Ich möchte Ihnen nun einige Fragen zu Ihrer Stimmung stellen. Während des letzten Monats, gab es da eine Zeit, in der Sie sich fast jeden Tag die meiste Zeit des Tages depressiv oder niedergeschlagen fühlten? Oder hat irgendjemand gesagt, dass Sie niedergeschlagen oder depressiv aussehen?

Patient: Also… weiß nicht genau. Im letzten Monat war's so, naja, mehr oder weniger fast jeden Tag. Hab mich durchweg schlapp und bedrückt gefühlt. Kaum mal 'nen Moment, wo's nicht gezogen hat. Meine Tochter hat auch gesagt, ich seh aus, als würd ich durchhängen, und 'ne Nachbarin hat gefragt, ob was nicht stimmt… Aber ich hab nur… so genickt und bin weggegangen.

Arzt: Bitte beschreiben Sie das genauer! Wie lange hielt dies an? 2 Wochen lang?

Patient: Naja… also ich denk, so richtig heftig war's locker drei Wochen… zwei Wochen vielleicht am schlimmsten, aber dann hat's nicht wirklich aufgehört, eher weitergezogen… so in etwa… mehr oder weniger… ständig halt.

[…]



Methods[1]

Step I: Get the „ground truth“ by human experts

Arzt: Ich möchte Ihnen nun einige Fragen zu Ihrer Stimmung stellen. Während des letzten Monats, gab es da eine Zeit, in der Sie sich fast jeden Tag die meiste Zeit des Tages depressiv oder niedergeschlagen fühlten? Oder hat irgendjemand gesagt, dass Sie niedergeschlagen oder depressiv aussehen?

Patient: Also… weiß nicht genau. Im letzten Monat war’s so, naja, mehr oder weniger fast jeden Tag. Hab mich durchweg schlapp und bedrückt gefühlt. Kaum mal ’nen Moment, wo’s nicht gezogen hat. Meine Tochter hat auch gesagt, ich seh aus, als würd ich durchhängen, und ’ne Nachbarin hat gefragt, ob was nicht stimmt… Aber ich hab nur… so genickt und bin weggegangen.

Please, highlight diagnosis-relevant text sections in the interview.

[1]So et al. (2024)

Methods[1]

Step I: Get the „ground truth“ by human experts

~~*Arzt*: Ich möchte Ihnen nun einige Fragen zu Ihrer Stimmung stellen. Während des letzten Monats, gab es da eine Zeit, in der Sie sich fast jeden Tag die meiste Zeit des Tages depressiv oder niedergeschlagen fühlten? Oder hat irgendjemand gesagt, dass Sie niedergeschlagen oder depressiv aussehen?~~

Patient: Also… weiß nicht genau. Im letzten Monat war’s so, naja, mehr oder weniger fast jeden Tag. Hab mich durchweg schlapp und bedrückt gefühlt. Kaum mal ’nen Moment, wo’s nicht gezogen hat. Meine Tochter hat auch gesagt, ich seh aus, als würd ich durchhängen, und ’ne Nachbarin hat gefragt, ob was nicht stimmt… Aber ich hab nur… so genickt und bin weggegangen.

- Depressive Stimmung, Müdigkeit, Keine manische oder hypomanische Episode

Please, highlight diagnosis-relevant text sections in the interview.
For each text section add the diagnosis-relevant symptoms (DSM-V) described there.

[1]So et al. (2024)

Methods$_1$

Step II: Get LLM to perform the same task as the human experts.

Arzt: Ich möchte Ihnen nun einige Fragen zu Ihrer Stimmung stellen. Während des letzten Monats, gab es da eine Zeit, in der Sie sich fast jeden Tag die meiste Zeit des Tages depressiv oder niedergeschlagen fühlten? Oder hat irgendjemand gesagt, dass Sie niedergeschlagen oder depressiv aussehen?

Patient: Also… weiß nicht genau. Im letzten Monat war's so, naja, mehr oder weniger fast jeden Tag. Hab mich durchweg schlapp und bedrückt gefühlt. Kaum mal 'nen Moment, wo's nicht gezogen hat. Meine Tochter hat auch gesagt, ich seh aus, als würd ich durchhängen, und 'ne Nachbarin hat gefragt, ob was nicht stimmt… Aber ich hab nur… so genickt und bin weggegangen.

System: Sie werden ein Interview erhalten. Achten Sie bei der Beantwortung der psychiatrischen Symptome im Zusammenhang mit Bipolarität oder Depression und dem entsprechenden Abschnitt darauf, in der Form [{Section': '...', ‚Symptoms': '...'}, {'Section ': '...', Symptoms ': '...'}, ...] zu antworten.\Wenn Sie der Meinung sind, dass es in einem bestimmten Abschnitt mehrere Symptome gibt, können Sie in der Form [{Section': '...', ‚Symptoms ': ['...' , '...']}] antworten.\Wenn in einem bestimmten Interview keine psychiatrischen Symptome im Zusammenhang mit Bipolarität und Depression auftreten, antworten Sie bitte in der Form [{Section': 'none', ‚Symptoms': 'none'}].\ Die folgenden 40 Symptome sollen berücksichtigt werden:\
Für eine schwere depressive Episode:Schlaflosigkeit, Hypersomnie, Psychomotorische Unruhe[…]

User: Wenn Sie der Meinung sind, dass die folgende Befragung psychiatrisch bedeutsame Symptome aufweist, geben Sie bitte an, um welche Symptome es sich handelt und in welchen Abschnitten des Textes Sie diese Symptome finden.

$_1$So et al. (2024)

Methods[1]

Step II: Get LLM to perform the same task as the human experts.

Statement	Section	Symptoms	Estimated Section	Estimated Symptoms
Arzt: Ich möchte Ihnen nun einige Fragen zu Ihrer Stimmung […]	Im letzten Monat war's so, naja, mehr oder weniger fast jeden Tag. Hab mich durchweg schlapp und bedrückt gefühlt.	Depressive Stimmung, Müdigkeit, Keine manische oder hypomanische Episode	Also… weiß nicht genau. Im letzten Monat war's so, naja, mehr oder weniger fast jeden Tag. Hab mich durchweg schlapp und bedrückt gefühlt. Kaum mal 'nen Moment, wo's nicht gezogen hat. Meine Tochter hat auch gesagt, ich seh aus, als würd ich durchhängen, und 'ne Nachbarin hat gefragt, ob was nicht stimmt	Depressive Stimmung, Signifikante Notlage

[1]So et al. (2024)

Methods[1]

Step III: Compare the ground truth section with the estimated section

Ground-truth section mid-token: 14

Patient: Also… weiß nicht genau. Im letzten Monat war's so, naja, mehr oder weniger fast jeden Tag. Hab mich durchweg schlapp und bedrückt gefühlt. Kaum mal 'nen Moment, wo's nicht gezogen hat. Meine Tochter hat auch gesagt, ich seh aus, als würd ich durchhängen, und 'ne Nachbarin hat gefragt, ob was nicht stimmt… Aber ich hab nur… so genickt und bin weggegangen.

Estimated section mid-token: 19

Patient: Also… weiß nicht genau. Im letzten Monat war's so, naja, mehr oder weniger fast jeden Tag. Hab mich durchweg schlapp und bedrückt gefühlt. Kaum mal 'nen Moment, wo's nicht gezogen hat. Meine Tochter hat auch gesagt, ich seh aus, als würd ich durchhängen, und 'ne Nachbarin hat gefragt, ob was nicht stimmt… Aber ich hab nur… so genickt und bin weggegangen.

→ Mid-token distance d = |14-19| = 5

[1]So et al. (2024)

Methods[1]

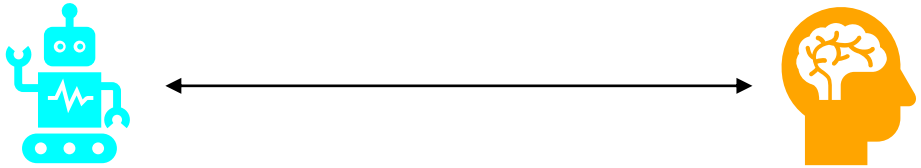
Step IV: Compare the ground truth symptoms with the estimated symptoms

Statement	Section	Matched Estimated Section	Symptoms	Estimated Symptoms	Accuracy
Arzt: Ich möchte Ihnen nun einige Fragen zu Ihrer Stimmung […]	Im letzten Monat war's so, naja, mehr oder weniger fast jeden Tag. Hab mich durchweg schlapp und bedrückt gefühlt.	Also… weiß nicht genau. Im letzten Monat war's so, naja, mehr oder weniger fast jeden Tag. Hab mich durchweg schlapp und bedrückt gefühlt. Kaum mal 'nen Moment, wo's nicht gezogen hat. Meine Tochter hat auch gesagt, ich seh aus, als würd ich durchhängen, und 'ne Nachbarin hat gefragt, ob was nicht stimmt	Depressive Stimmung, Müdigkeit, Keine man./hypoman. Epis -	Depressive Stimmung, - - Signifikante Notlage	1 0 0 0

[1]So et al. (2024)

Methods

Step V: Evaluate the LLM performance with the average human performance

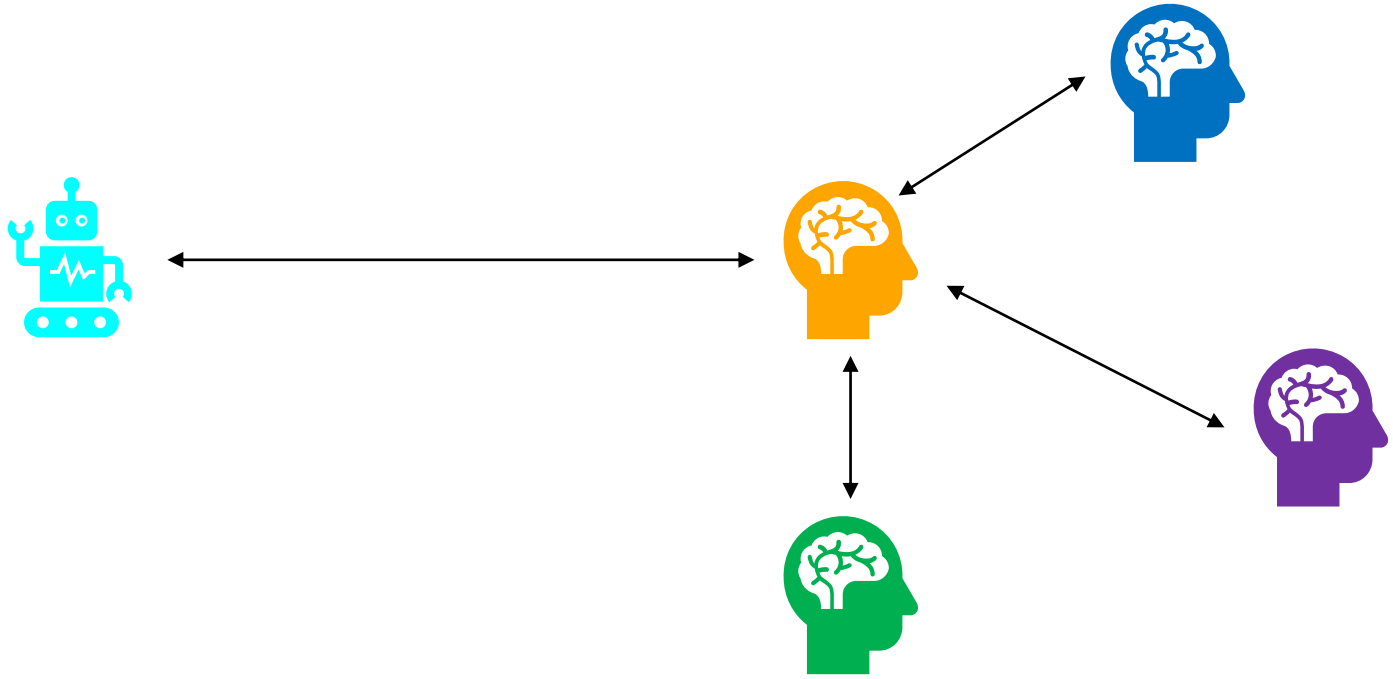


Is that helpful ?

Model	Accuracy	Pos. Predicitve value	Neg. predictive value	Recall	F1-score
GPT-4 Turbo	.71	.72	.95	.75	.72

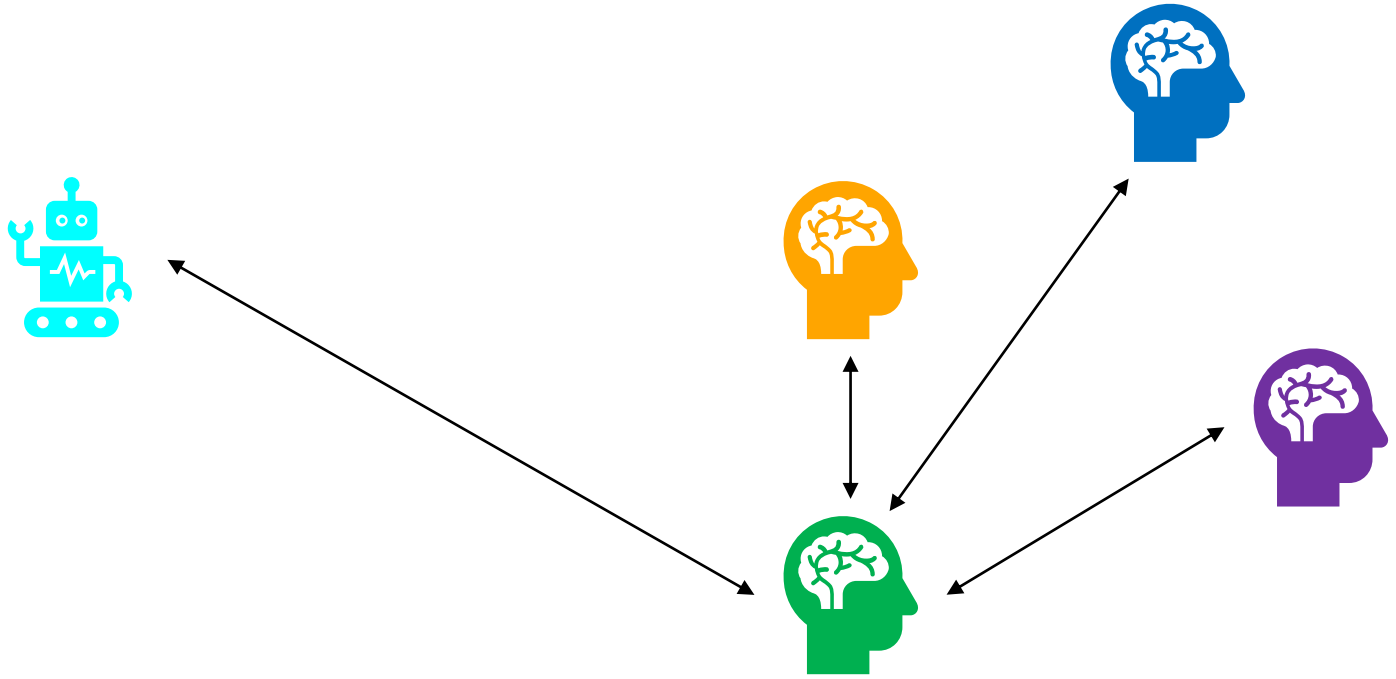
Methods

Step V: Evaluate the LLM performance with the average human performance

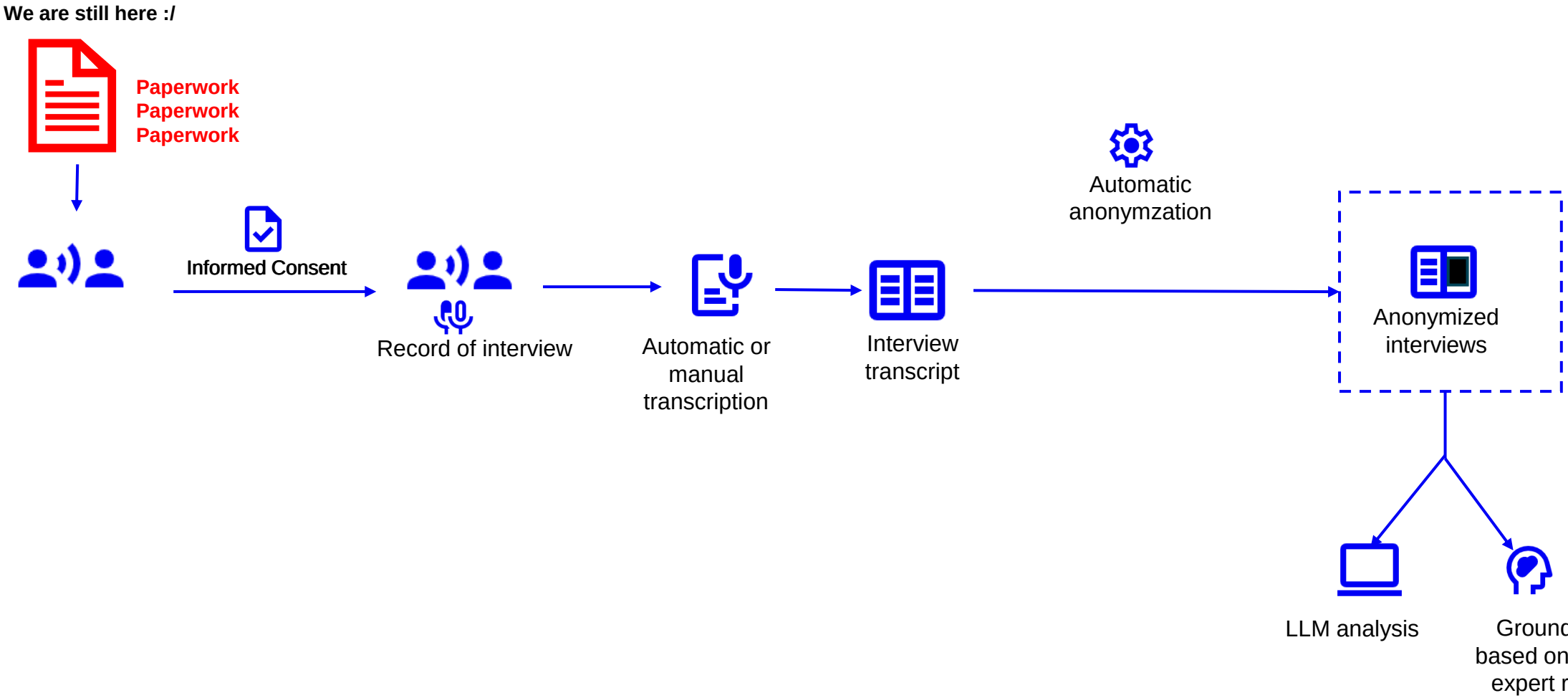


Methods

Step V: Evaluate the LLM performance with the average human performance



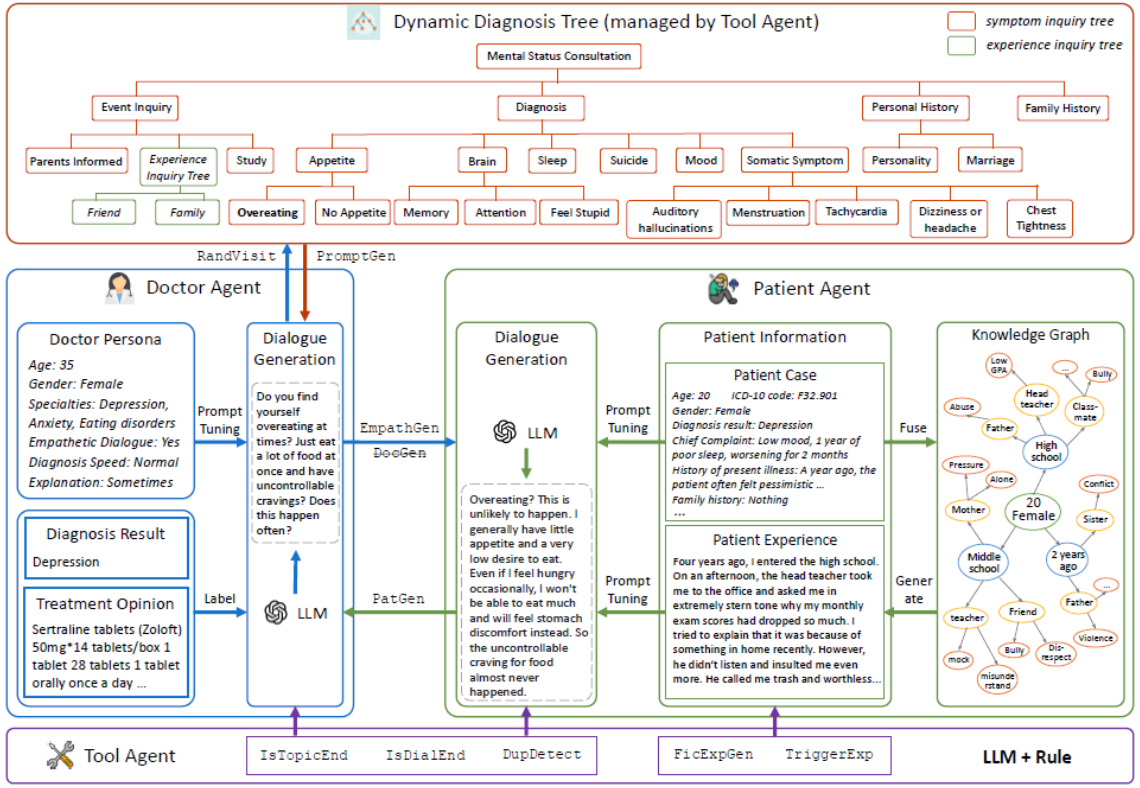
Challenge of clinical patient data



20

Opportunity for synthetic interview data?

Master thesis Project of Esther Stumm



Yin et al. (preprint)

Open questions?
Opportunities for cooperations!

Cases where human textwork processes were supported/substituted by LLM-Agents?

Relevant measures for a possible implementation of an AI system?

Experience/knowledge about the creation of synthetic interview data?