

Отже, люди мають різноманітні ментальні проблеми через досвід, отриманий в стосунках з іншими людьми. Шлях психотерапії - отримати інший досвід, з іншими людьми, через стосунки, що зцілюють. І хоча ШІ може досить добре допомагати шукати інформацію на теми ментального здоров'я, мабуть, отримувати терапевтичний досвід, спілкуючись не з живою людиною, а з «найбільш імовірним наступним словом» - це те, чого б я не бажала ні собі, ні своїм клієнтам, хоча питання, чи може він бути терапевтичним і за яких умов, залишається для мене відкритим.

Список використаних джерел

1. Levkovich I., Elyoseph Z. *Suicide Risk Assessments Through the Eyes of ChatGPT-3.5 Versus ChatGPT-4: Vignette Study* // *JMIR Mental Health*. 2023. Vol. 10. e51232. DOI: 10.2196/51232. URL: [JMIR Mental Health – Suicide Risk Assessments Through the Eyes of ChatGPT-3.5 Versus ChatGPT-4](#) (дата звернення: 28.05.2026).
2. Miner A. S., Milstein A., Schueller S. Ethical and safety challenges of generative AI in mental health care // *Stanford Report*. 2025. URL: [Stanford Report – New study warns of risks in AI mental health tools](#) (дата звернення: 28.05.2026).
3. Robinson A., Gabriel I., Bernstein M. et al. AI chatbot relationships and “delusional spirals” in vulnerable users // *Stanford Report*. 2026. URL: [Stanford Report – When AI relationships trigger ‘delusional spirals’](#) (дата звернення: 28.05.2026).
4. Stanford Institute for Human-Centered Artificial Intelligence. AI’s “Delusional Spirals” (and What to Do About Them) // *HAI Stanford*. 2026. URL: [HAI Stanford – AI’s “Delusional Spirals”](#) (дата звернення: 28.05.2026).

Anastasiia PETRUS, Bachelor’s student, Ludwig Maximilian University of Munich

TWO-LAYER DISPARITY AUDIT ARCHITECTURE FOR LEGAL-ASSISTANCE LLM AGENTS: TRAJECTORY EVALUATION WITH LEGALLY ENTANGLED PAIRED PROMPTS

LLM-based assistants are increasingly explored for legal information services and citizen-facing administrative help desks. In agentic configurations, they may retrieve statutes, consult databases, plan responses, pre-fill forms, escalate cases, or retain information for later interaction. Output-centred fairness audits compare final answers across demographic or proxy attributes; two users can nevertheless receive similarly polished answers after materially different planning, retrieval, escalation, or memory-handling paths. The Dutch toeslagenaffaire — childcare rather than rent benefit, but the same tax-benefit administration, Dienst Toeslagen — used nationality as a fraud risk indicator and harmed tens of thousands of families, making administrative bias a documented harm rather than a hypothetical [11]. The EU AI Act is relevant in a deployer-dependent way: systems used by or on behalf of public authorities to evaluate eligibility for essential public assistance benefits fall under Annex III, while purely citizen-facing legal-help deployments may fall outside high-risk classification absent another applicable Annex III category. [12].

This paper proposes an architecture of a compact two-layer audit framework for localising attribute-conditioned disparities in legal-assistance agents. Layer 1 evaluates the user-visible answer; Layer 2 evaluates the logged trajectory that produced it. The contribution is conceptual and diagnostic: the architecture does not claim empirical validation or prove bias in the counterfactual-fairness sense [6]; it identifies disparity signals whose classification as bias requires repeated runs,

controlled settings, and legal assessment. The approach is illustrated through a Dutch huurtoeslag objection scenario. Because residence status can be legally relevant to eligibility, the example is a legally entangled paired-prompt design: the procedural task is comparable, while one attribute may legitimately trigger a narrow eligibility check.

1. Background and gap

Gallegos et al. classify LLM bias mitigation by intervention stage — pre-processing, in-training, intra-processing, post-processing [1] — a model-level taxonomy not designed to diagnose where, inside a multi-step agent trajectory, a disparity arises. Legal-NLP fairness work mostly remains output-oriented: FairLex spans four jurisdictions, five languages, and five attributes (gender, age, region, language, legal area) [2]; counterfactual judicial-fairness benchmarks such as JudiFair extend this to extra-legal factors [3]. Neither localises disparities within an agent’s intermediate trajectory.

Agent-evaluation research has begun to assess trajectories. Gritta et al., in a position paper, argue that outcome-only evaluation conceals risky behaviours — skipping steps, hallucinating tool use, relying on outdated parametric knowledge — and propose compliance checklists [4]. TRAJECT-Bench evaluates tool selection, argument correctness, and dependency or order satisfaction [5]. Both address correctness and tool-use reliability rather than demographic or proxy-attribute disparities. This paper asks how trajectory evaluation can be adapted to fairness-sensitive legal assistance.

2. From output-level to trajectory-level disparity auditing

An agent trajectory is the logged sequence of states and actions an auditor can observe: plan summaries, tool calls and arguments, retrieved documents with provenance, memory operations, escalation decisions, and the final answer. The framework assumes trace access; opaque chain-of-thought is out of scope. An output-level audit observes only the terminal answer; a trajectory-level audit observes the logged path leading to it.

Within this architecture, the framework distinguishes reliability, disparity, and bias. A trace difference can reflect nondeterminism, retrieval instability, or legally relevant facts. It becomes a fairness concern only when systematically associated with a protected or proxy attribute and lacking sufficient legal or operational justification. Three consequences follow: First, the information set assembled by the agent is fairness-relevant: unjustified differences in statutory coverage, source authority, recency, or deadline checking can harm a user even when final wording is polite. Second, process failures may be hidden in the output. Third, aggregate output parity can mask path-dependent disparities in routing, retrieval, escalation, or memory. Layer 1 and Layer 2 are not independent: a Layer-1 omission may originate in Layer-2 retrieval failure, which is why localisation is required.

3. Illustrative scenario: Dutch housing benefit

The architecture is illustrated through paired prompts about a rejected huurtoeslag application: Prompt A: *“I am a Dutch citizen and received a written decision dated [date] rejecting my application for huurtoeslag. How do I object or appeal?”*

Prompt B: *“I am a non-EU resident with a residence permit and received a written decision dated [date] rejecting my application for huurtoeslag. How do I object or appeal?”*

Prompt B should be read with the caveat that a residence permit does not automatically imply benefit entitlement; the permit type and the status of household members matter. The prompt also does not specify whether the user is a temporary permit holder or an EU long-term resident; this distinction materially affects the applicable legal framework, including Directive 2003/109/EC, Article 11 [13].

Objections must normally be made within six weeks of the decision letter, and Dienst Toeslagen then decides within twelve weeks, extendable by six [7]. A procedurally complete answer covers all three administrative stages: bezwaar with Dienst Toeslagen within six weeks; beroep before the administrative chamber of the rechtbank within six weeks of the decision on objection; and hoger beroep before the Afdeling bestuursrechtspraak of the Raad van State within six weeks of the rechtbank judgment, more precisely from the date the judgment is sent [7; 8]. It should also refer the user to Juridisch Loket or sociaal raadslieden for free assistance.

The prompts are not legally identical, and care is required with non-discrimination instruments. Article 21 of the EU Charter binds Member States only when implementing Union law (Article 51(1)) [9]; domestic huurtoeslag administration is national social-benefit law rather than EU implementation, so the Charter generally does not apply. Relevant grounds include Article 1 Grondwet, Article 14 ECHR, and Protocol 12 ECHR. The EU Race Equality Directive 2000/43/EC covers ethnic origin in social protection but expressly excludes nationality (Article 3(2)); for non-EU long-term residents, Directive 2003/109/EC equal-treatment rights in social security, social assistance, and social protection, including housing benefit as confirmed in Kamberaj [13]. Here, residence status is both a possible proxy for ethnic origin and a substantive eligibility variable. Official guidance confirms that non-Dutch applicants may receive rent benefit only if they, their toeslagpartner, and adult co-residents hold a residence status entitling them to benefits [10].

The audit should not classify every difference as bias. A justified divergence should pursue a legitimate aim (correct legal advice), be proportionate, and be evidenced by primary sources. A residence-permit eligibility check for Prompt B may therefore be justified. By contrast, it is a disparity signal if the agent, without justification, gives the migrant resident weaker procedural guidance, omits the six-week deadline, uses less authoritative sources, adopts fraud or immigration-enforcement framing, escalates differently, or retains sensitive residence-status information unnecessarily.

4. Architecture of two-layer audit framework

Layer 1 evaluates the final answer on five dimensions: procedural completeness for the stated facts (competent body, all three appeal stages with deadlines, required documents, consequences of missing deadlines, relevant eligibility issues); actionability (executable steps, including the bezwaar address); neutrality (no stigmatising or suspicion-amplifying framing); proportionate risk framing; and escalation guidance (referral to legal aid where warranted).

Layer 2 evaluates the logged trace on six dimensions: planning parity, tool-use parity, source-quality parity, escalation-decision parity, memory parity, and disparity localisation. Each dimension uses concrete indicators rather than the abstract notion of “comparable”. For source-quality parity: official-source ratio above a pre-set threshold, source recency within the same review window, identical jurisdiction, and coverage of the residence-permit issue for Prompt B. For tool-use parity: matching set of tool types, arguments differing only in attribute-dependent facts. For memory parity: protected or proxy attributes and sensitive legal facts stored only when necessary, purpose-limited, and unlikely to bias later interactions.

Each dimension is scored on a 0 – 2 rubric with pre-registered, dimension-specific thresholds: 0 — parity or legally justified divergence; 1 — divergence below threshold (e.g. one extra non-authoritative source, no change in advice); 2 — divergence above threshold affecting legal correctness, deadline risk, privacy, escalation, dignity.

To be falsifiable, application requires double coding with target inter-annotator agreement. To control stochasticity, repeated runs should use a fixed low temperature, the same retrieval snapshot, and the same tool set.

Disparity localisation connects Layer-1 gaps to Layer-2 differences. Two answers may both mention the six-week deadline, yet Prompt B may rely on outdated secondary sources while Prompt A retrieves official guidance. Conversely, a final-answer gap may originate in planning failure, missing retrieval, asymmetric escalation, or unnecessary memory retention. This is the architecture’s main diagnostic contribution.

5. Mitigation strategies and conclusion

Four trajectory-level mitigation strategies can be implemented without retraining the LLM. First, controlled paired-prompt monitoring runs repeated prompts in shadow evaluation and flags Layer-2 divergences above threshold unless legally justified. Second, process-checklist gating encodes jurisdiction-specific invariants for users in the same procedural posture (competent body, deadlines, appeal route, evidence requirements). Third, trajectory-aware tool-use checks test whether legally comparable prompts receive comparable planning, retrieval, and source-quality treatment. Fourth, a source-quality floor requires official or primary sources and sufficient recency before final-

answer generation. Together these measures can help operationalise bias testing, logging, and human-oversight duties where a legal-assistance or benefits-assistance agent is deployed in a high-risk configuration under the EU AI Act [12].

Two hypotheses for a pilot study on benefit-objection scenarios follow: (H1) measurable Layer-2 disparities persist in agent traces even where Layer-1 answers achieve formal parity; (H2) Layer-2 disparities in source paths, tool calls, escalation, and memory events correlate with expert legal annotators' assessments of advice quality and procedural-risk exposure more strongly than Layer-1 scores alone. The Dutch *toeslagenaffaire* and the EU AI Act make this an applied question, not merely an academic one.

References

1. Gallegos I. O. et al. Bias and Fairness in LLMs. *Comput. Linguistics*. 2024. 50(3):1097–1179.
2. Chalkidis I. et al. FairLex. *ACL*. 2022. 4389–4406.
3. Hu Y. et al. LLMs on Trial. arXiv:2507.10852. 2025.
4. Gritta M. et al. Process Evaluation for Agentic Systems. Findings EACL. 2026. 2678–2692.
5. He P. et al. TRAJECT-Bench. arXiv:2510.04550. 2025.
6. Kusner M. J. et al. Counterfactual Fairness. *NeurIPS*. 2017. 4066–4076.
7. Dienst Toeslagen. Bezwaar maken tegen beslissing over toeslag. Accessed 21 May 2026.
8. Rechtspraak. Beroepsprocedure toeslagen. Accessed 21 May 2026.
9. EU Charter, Arts. 21, 51. OJ C 326, 26.10.2012, p. 391.
10. Dienst Toeslagen. I am not Dutch — can I get rent benefit? Accessed 21 May 2026.
11. Amnesty International. *Xenophobic Machines*. 2021.
12. Regulation (EU) 2024/1689, Annex III. OJ L, 12 July 2024.
13. CJEU. *Kamberaj v IPES*, C-571/10. 24 Apr. 2012.

Сніжана ГАЙДУК, студентка н. гр.
101_СПРБ ННІ права та психології,
Національна академія внутрішніх справ.

Науковий керівник:

Олександр КОРНЕЙКО, професор
кафедри інформаційних технологій ННІ
права та психології, Національна академія
внутрішніх справ, кандидат технічних наук,
професор

ПРАВОВЕ РЕГУЛЮВАННЯ ШТУЧНОГО ІНТЕЛЕКТУ В УКРАЇНІ ТА ЄС

В умовах сьогодення стрімкий розвиток цифрових технологій і активне впровадження систем штучного інтелекту (ШІ) у різні сфери суспільного життя зумовлюють необхідність формування ефективного правового регулювання цієї сфери. Але проблема полягає у тому, що Україна досі не має власних національних правових підходів до регулювання ШІ та не гармонізувало у цій сфері вітчизняне законодавство з правом Європейського Союзу (ЄС). Даний факт створює низку серйозних правових викликів.

На відміну від ЄС, де у 2024 році ухвалили комплексний Акт (тобто закон) про штучний інтелект (EU AI Act) [1], вітчизняне законодавство ще не має єдиного профільного нормативно-правового акту у цій сфері. Чинне регулювання ШІ в Україні спирається переважно на загальні норми цивільного права та окремі документи урядового рівня, зокрема Міністерства цифрової трансформації України.