UNIVERSITAS
GADJAH MADA
**Grounding Case-Based Reasoning (CBR) untuk Mitigasi Halusinasi Large Language Model (LLM) pada
Sistem Pendukung Keputusan Klinis Polycystic Ovary Syndrome (PCOS) Berbasis Kriteria Rotterdam**
Diaz Astari Beningati, Prof. Dr.Eng. Ir. Igi Ardiyanto, S.T., M.Eng., IPM., ASEAN Eng., SMIEEE.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

INTISARI

Large Language Model (LLM) berpotensi mendukung keputusan klinis Polycystic Ovary Syndrome (PCOS) melalui penalaran naratif yang dapat dijelaskan, tetapi rentan menghasilkan halusinasi intrinsik, yaitu keluaran yang bertentangan dengan data pasien dan kriteria diagnosis yang diberikan. Penelitian ini mengevaluasi apakah *grounding* berbasis *Case-Based Reasoning* (CBR) dapat menurunkan halusinasi intrinsik tersebut pada sistem pendukung keputusan klinis berbasis Kriteria Rotterdam. Basis kasus disusun dari 541 kasus dataset Kottarathil (2020) yang diperkaya dengan status kriteria Rotterdam, kondisi *mimic*, status diagnosis, fenotipe, dan faktor risiko melalui satu modul kriteria operasional berbasis Kriteria Rotterdam dan pedoman Teede et al. (2023). Evaluasi dilakukan dalam dua tahap. Evaluasi pertama membandingkan *Hierarchical Weighted Heterogeneous Similarity Measure* (HWHSM) berbasis *Clinical Decision Tree* (CDT) dengan HEOM melalui *Leave-One-Out Cross-Validation*. Evaluasi kedua membandingkan gpt-4o-mini tanpa dan dengan lima kasus serupa hasil *retrieval* HWHSM, dengan *prompt* identik yang memuat kriteria secara eksplisit. Setiap keluaran diuraikan menjadi 13 proposisi dan diverifikasi secara deterministik terhadap modul kriteria, lalu diuji dengan uji McNemar dan koreksi Holm. Pada Evaluasi pertama, seleksi atribut berbasis CDT meningkatkan akurasi *reuse* fenotipe dari 85,95% menjadi 97,78%, sedangkan pembobotan hierarkis tidak memberikan peningkatan tambahan yang berarti. Pada evaluasi kedua, halusinasi intrinsik keseluruhan turun dari 40,48% menjadi 39,00% tanpa perbedaan yang signifikan, tetapi turun signifikan pada penilaian *mimic* dan status diagnosis dari 26,99% menjadi 23,11% (p Holm = 0,023) pada evaluasi utama. Sumber halusinasi terbesar pada kedua arsitektur adalah kegagalan menerapkan aturan dua dari tiga kriteria Rotterdam, yang terjadi pada 57–60% pasien dengan tepat dua kriteria terpenuhi meskipun penilaian per kriteria oleh model sudah benar. Uji keterulangan pada tiga run menunjukkan simpangan baku halusinasi di bawah 1 poin persentase, dengan arah penurunan pada penilaian *mimic* dan status diagnosis yang konsisten tetapi signifikan hanya pada run utama, sedangkan penyajian data pasien sebagai narasi tidak mengubah tingkat halusinasi secara signifikan. Temuan ini menunjukkan bahwa *grounding* berbasis CBR cenderung meningkatkan kesetiaan LLM terhadap penilaian kondisi mimic dan status diagnosis, tetapi tidak pada penerapan aturan dua dari tiga kriteria, sehingga keluaran sistem diposisikan sebagai rekomendasi berbasis kriteria yang mendukung keputusan klinisi.

Kata kunci : Polycystic Ovary Syndrome (PCOS), Case-Based Reasoning (CBR), Large Language Model (LLM), halusinasi intrinsik, *faithfulness*, gpt-4o-mini, Clinical Decision Tree (CDT), Hierarchical Weighted Heterogeneous Similarity Measure (HWHSM), Kriteria Rotterdam.

UNIVERSITAS
GADJAH MADA
**Grounding Case-Based Reasoning (CBR) untuk Mitigasi Halusinasi Large Language Model (LLM) pada
Sistem Pendukung Keputusan Klinis Polycystic Ovary Syndrome (PCOS) Berbasis Kriteria Rotterdam**
Diaz Astari Beningati, Prof. Dr.Eng. Ir. Igi Ardiyanto, S.T., M.Eng., IPM., ASEAN Eng., SMIEEE.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

ABSTRACT

Large Language Models (LLMs) can support clinical decision making for Polycystic Ovary Syndrome (PCOS) through explainable narrative reasoning, but they are prone to intrinsic hallucination, that is, output that contradicts the patient data and diagnostic criteria provided. This study evaluates whether grounding based on Case-Based Reasoning (CBR) reduces such hallucination. The case base was built from 541 cases of the Kottarathil (2020) dataset and enriched with Rotterdam criteria status, mimicking conditions, diagnostic status, phenotype, and risk factors through a single operational criteria module based on the Rotterdam Criteria and the guideline of Teede et al. (2023). Evaluation was performed at two levels. The first evaluation compared a Clinical Decision Tree (CDT) based Hierarchical Weighted Heterogeneous Similarity Measure (HWHSM) with HEOM using Leave-One-Out Cross-Validation. The second level compared gpt-4o-mini without and with the five most similar cases retrieved by HWHSM, using identical prompts that stated the criteria explicitly. Each output was decomposed into 13 propositions, verified deterministically against the criteria module, and tested with McNemar's test and Holm correction. At the first evaluation, CDT-based attribute selection increased phenotype reuse accuracy from 85.95% to 97.78%, whereas hierarchical weighting added no meaningful improvement. At the second evaluation, overall intrinsic hallucination decreased from 40.48% to 39.00% without a significant difference, but CBR grounding significantly reduced hallucination in the assessment of mimicking conditions and diagnostic status from 26.99% to 23.11% (Holm-adjusted p = 0.023) in the main run. The largest source of hallucinations in both architectures is the failure to apply the two out of three Rotterdam criteria rule, which occurs in 57–60% of patients with exactly two criteria met, even though the model's assessment per criterion is correct. A repeatability test over three runs showed standard deviations of hallucination below 1 percentage point, with a consistent reduction for mimicking conditions and diagnostic status that was significant only in the main run, while presenting patient data as narrative text did not significantly change hallucination. These findings indicate that CBR-based grounding tends to improve the fidelity of LLM to the assessment of mimic conditions and diagnostic status, but not in the application of the two out of three criteria rule; the system output is therefore positioned as a criteria-based recommendation that supports clinicians.

Keywords : Polycystic Ovary Syndrome (PCOS), Case-Based Reasoning (CBR), Large Language Model (LLM), intrinsic hallucination, faithfulness, gpt-4o-mini, Clinical Decision Tree (CDT), Hierarchical Weighted Heterogeneous Similarity Measure (HWHSM), Rotterdam Criteria.