International Journal of Technology Assessment in Health Care

www.cambridge.org/thc

Method

Cite this article: González-González E, Castilla-Rodríguez I, Dorta-Hernández JA (2026). Ontology-driven generation of parameters for health technology assessment models: a prompt engineering study. *International Journal of Technology Assessment in Health Care*, **42**(1), e47, 1–10
https://doi.org/10.1017/S0266462326103754

Received: 13 November 2025
Revised: 26 March 2026
Accepted: 04 April 2026

Keywords:
Ontology Population; Generative Artificial Intelligence, Health Technology Assessment; Economic Evaluation Model; Conceptual Model; Prompt Engineering; Knowledge Synthesis; Automation

Corresponding author:
Iván Castilla Rodríguez,
Email: icasrod@ull.edu.es

E. G. G., I. C. R. and J. A. D. H. these authors are contributed equally.

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike licence (http://creativecommons.org/licenses/by-nc-sa/4.0), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the same Creative Commons licence is used to distribute the re-used or adapted article and the original article is properly cited. The written permission of Cambridge University Press or the rights holder(s) must be obtained prior to any commercial use.

Ontology-driven generation of parameters for health technology assessment models: a prompt engineering study

Evelio González-González, Iván Castilla-Rodríguez and Joel Aday Dorta-Hernández

Departamento de Ingeniería Informática y de Sistemas, Universidad de La Laguna, Spain

Abstract

Objectives: Ontologies support transparent and reproducible conceptual modeling in Health Technology Assessment (HTA), but their population remains resource-intensive and reliant on expert input. This study evaluates the feasibility, reliability, and methodological implications of using generative artificial intelligence (GenAI) to populate ontology individuals for HTA applications.

Methods: A factorial experimental framework was developed using the Ontology for Simulation Modeling (OSDi) and three HTA-relevant use cases of varying complexity. Two GenAI systems were evaluated under multiple experimental conditions, including prompting strategy, serialization format, and provision of supporting information. Generated ontology individuals were validated by an HTA expert and assessed across four quality dimensions: consistency, relevance, completeness, and adequacy. Multivariate and regression analyses were conducted to examine the effects of experimental factors on quality outcomes and hallucination likelihood.

Results: GenAI systems successfully generated ontology individuals across use cases, although performance varied by quality dimension and experimental condition. Iterative prompting significantly improved completeness, while serialization format strongly influenced reliability, with Turtle serialization associated with substantially lower hallucination likelihood compared with XML. Other factors showed dimension-specific effects, highlighting the multidimensional nature of ontology quality. Errors occurred more frequently in structurally complex ontology components, suggesting a relationship between ontological complexity and generative performance.

Conclusions: GenAI-assisted ontology population can enhance the efficiency and scalability of HTA conceptual modeling, enhancing the agility of HTA agencies in exploratory phases. Its effective use requires structured prompting, appropriate representation formats, and expert validation. Further research should evaluate its impact on HTA decision modeling workflows and validation frameworks.

Introduction

Economic evaluation models are key tools for health technology assessment (HTA) (1), as they support policy-makers and HTA practitioners in comparing the costs and health outcomes of alternative interventions. For methodological rigor and relevance, such models require a coherent integration of clinical, epidemiological, and economic knowledge. Traditionally, building these models has been resource-intensive, often requiring substantial time and specialized expertise, due to extensive data gathering and expert interpretation.

Ontologies, defined as formal representations of domain knowledge, have been proposed as a solution (2), providing a semantic framework to define entities such as diseases, interventions, costs or outcomes, and their relationships (3;4). This structure enhances consistency, interoperability, and reusability, allowing HTA practitioners to standardize model conceptualization across evaluation contexts.

In parallel, generative artificial intelligence (GenAI), particularly large language models (LLMs), has transformed knowledge creation and management (5). LLMs efficiently summarize, extract, and generate domain-specific content, thereby reducing manual effort in evidence synthesis, automating processes that previously required substantial expertise (6).

This article explores how combining ontologies with GenAI enables scalable and automated ontology population. Ontologies are conceived not as static repositories but as dynamic guides for LLMs, whose outputs are constrained by semantic rules to generate relevant, context-aware individuals. This addresses a critical bottleneck in HTA modeling: the costly and error-prone manual curation of domain-specific evidence.

Accelerating the synthesis of accurate clinical and economic data requires expert review of literature and guidelines, creating a gap that limits scalability. GenAI's natural language

capabilities can mitigate this by automatically extracting and synthesizing knowledge. Yet controlling LLM behavior remains difficult due to their stochastic nature and the risk of hallucinations, defined here as the generation of content that is syntactically plausible but factually incorrect, unsupported by evidence, or inconsistent with the target knowledge base.

Prompt engineering has improved control over LLM outputs through structured prompts and iterative refinement (7). However, its use for ontology-guided individual generation is still underexplored. Most GenAI healthcare research focuses on text generation and decision support (8;9), whereas robust methodological frameworks for systematic model population remain scarce (10;11), limiting reproducibility and reliability in regulated HTA domains.

This study proposes combining the expressive power of ontologies with the generative capacity of LLMs, steered by prompt engineering, to streamline the development of evaluation models. We hypothesize that ontologies can structure and constrain LLM output, producing valid, semantically rich individuals aligned with the target ontology, Ontology for Simulation of Diseases (OSDi) (12). An experimental framework was designed to generate individuals via ontology-driven prompts, testing strategies for syntactic validity, semantic completeness, and practical utility in accelerating health economic evidence synthesis.

The remainder of the paper is organized as follows: Section "State of the art" reviews prior work on ontology-based modeling, GenAI, and prompt engineering; Section "The OSDi ontology" introduces the OSDi ontology; Section "Methodology" details the methodology and evaluation criteria; Section "Results and discussion" presents and discusses results; and Section "Conclusions and further work" concludes with implications and future directions.

State of the art

GenAI systems (exemplified by ChatGPT-4, Claude, Perplexity, and Mixtral) are transforming healthcare by automating data integration and analysis (13). This enables advances in evidence synthesis, personalized medicine, and population health management (14;15). Current LLMs can efficiently extract terminologies, uncover relationships, and generate context-sensitive knowledge (16–18). In this sense, LLMs are increasingly being used in evidence screening and synthesis, allowing to automate health economic modeling (19). Indeed, ISPOR has established a dedicated group reflecting the growing importance of this topic (https://www.ispor.org/member-groups/task-forces/genai-for-heor-slrs-task-force).

Recent research on GenAI in healthcare focuses mainly on automated report generation, question answering, decision support, and evidence synthesis. Reviews highlight both the benefits (efficiency, automation, triage) and risks (hallucinations, bias, privacy breaches) of these tools. Chustecki (20) summarizes opportunities and regulatory challenges, whereas Panteli et al. (21) emphasize GenAI's role in public health surveillance and communication, noting concerns about equity, privacy, and governance.

In HTA, Fleurence et al. (22) and Reason et al. (23) offer insights on how GenAI supports literature reviews and evidence synthesis. Studies such as Qureshi et al. (24), Reason et al. (25), and Li et al. (26) show that LLMs can summarize and extract information efficiently but still require human validation. Gartlehner et al. (27) and Schopow et al. (28) report similar findings: high accuracy but a need for semi-automated workflows. Szabó et al. (29) explore data extraction from cost-effectiveness models in HTA reports, finding limited performance in complex modeling assumptions.

Prompt engineering emerges as a key skill for the reliable use of GenAI in medicine (30–32). Well-designed, iterative prompts can enhance domain specificity and compliance with medical standards. Recent frameworks, including programmatic and feedback-driven approaches, improve model consistency while balancing computational efficiency (18;31;32).

However, none of these studies address how LLMs can interact with formal semantic technologies. Ontologies structure knowledge through defined entities (diseases, interventions, costs, outcomes) and relationships, facilitating integration across domains. Frameworks like *SNOMED CT* (https://www.snomed.org/) demonstrate how ontologies enhance decision support, predictive analytics, and population health systems. Ambalavanan et al. (33) highlight challenges in scalability, governance, and privacy, suggesting AI-assisted ontology maintenance as a future direction. In health economics, automatically generating ontology-aligned individuals could streamline decision trees and simulations, promoting standardized evaluations among HTA agencies. Persistent challenges include ensuring semantic accuracy, reducing bias, and maintaining ethical oversight (34).

Emerging studies combine LLMs with ontologies for enrichment, text-to-OWL transformation, and alignment, showing potential but lacking methodological consistency. Ouédraogo et al. (10) propose an automated pipeline where ChatGPT-3 converts treatment guidelines for multidrug-resistant tuberculosis into OWL axioms, producing a richer ontology than a semi-automated baseline. Complementarily, prompt engineering research provides structured methodologies such as the Goal-Prompt-Evaluation-Iteration (GPEI) framework by Velásquez Henao et al. (35), which guides iterative prompt refinement through defined objectives, evaluation, and correction.

The OSDi ontology

The ontology used in this study is OSDi (12), an extension of the earlier RaDIOS ontology (2). OSDi provides a flexible semantic framework to represent knowledge about diverse diseases and healthcare interventions, organizing clinical and economic concepts into classes and relationships that capture disease characteristics, progression, and intervention effects. For its development from the earliest versions of the research, the authors adopted selected steps from the NeOn methodological framework (36), because it offers a set of flexible alternatives that can be tailored to the specific needs of the ontology developer. Among its proposed scenarios, "Scenario 1" which focuses on developing an ontology from scratch without relying on pre-existing knowledge resources aligns with the objectives of our work. The process in this case comprised the following main steps: (i) domain and scope definition, (ii) identification of core concepts and relationships, (iii) ontology formalization in OWL, (iv) validation and refinement with domain experts, and (v) ontology population with individuals. Section "Methodology" focuses in more detail on step (vi), namely the semi-automated creation of ontology individuals using LLMs, which constitutes a key contribution of this work.

At its core, OSDi models disease dynamics through classes such as *Development*, *Stage*, and *Manifestation*, allowing for the representation of disease presentation, evolution, and severity (e.g., distinguishing mild from severe developments or defining symptom probabilities at different stages). Disease progression is further

https://doi.org/10.1017/S0266462326103754 Published online by Cambridge University Press

detailed through *Pathways*, which describe alternative routes between stages or manifestations, conditioned by previous events, interventions, or probabilistic parameters.

The ontology also represents *Populations*, *Interventions*, and *Health resources* involved in HTA, enabling the formal specification of clinical follow-up and treatment processes.

Moreover, `OSDi` provides mechanisms to represent *Parameters* used in health economic modeling, including resources, distributions, rates, and risks. It incorporates utilities and disutilities, that is, quantitative measures of health-related quality of life, typically ranging from 0 (death) to 1 (perfect health), or the corresponding loss due to disease, adverse events, or side effects.

Compared with `RaDIOS`, `OSDi` adds greater detail to parameter characterization, distinguishing between *FirstOrder*, *SecondOrder*, and *DeterministicParameters*, and allowing *CalculatedParameters* defined by expressions. These enhancements enable not only the storage of numerical values but also the encoding of statistical properties and inferential rules, thereby supporting richer and more realistic simulation models.

Overall, `OSDi` standardizes the description of disease modeling components, from clinical presentation and progression to uncertainty parameterization and intervention effects. It is particularly suited to economic evaluation models, such as decision trees and discrete-event simulations, providing a reusable structure for health economics.

Currently, `OSDi` includes individuals for two diseases: profound biotinidase deficiency (PBD) and type 1 diabetes mellitus (T1DM). PBD is an autosomal recessive disorder of biotin metabolism manifesting in childhood, diagnosed via biochemical, enzymatic, and genetic tests, and treated with biotin supplementation (37). The corresponding individuals (https://w3id.org/ontologies-ULL/OSDi/individuals/PBD.ttl) replicate the PBD submodel from Vallejo-Torres et al. (38). T1DM, in contrast, is a chronic autoimmune disease marked by beta-cell destruction and absolute insulin deficiency (39), managed through continuous insulin therapy and glucose monitoring. Its individuals (https://w3id.org/ontologies-ULL/OSDi/individuals/T1DM.ttl) mirror the model by Castilla-Rodríguez et al. (40).

Methodology

The methodology combines factorial experimental design with prompt engineering techniques (35). A full-factorial setup was used to generate ontology individuals representing disease progression (pathogenesis, stages, manifestations) and their associated utilities and disutilities for PBD, forming the basis for estimating health indicators such as life expectancy and quality-adjusted life-years.

LLM interactions were conducted through the Perplexity AI platform (https://www.perplexity.ai/) under a Pro subscription, which provides controlled experimental "spaces" for uploading and querying documents. Both the Perplexity deep research model and ChatGPT-4.1 (https://chatgpt.com/) (41) were used. The platform's ability to cite sources proved valuable, especially when no external references were supplied. Each experiment included contextual information and explicit task instructions for the model.

AI-generated individuals were incorporated into separate ontology versions while maintaining the shared `OSDi` structure. Ontology management and editing were performed using Protégé (42) and Visual Studio Code (https://code.visualstudio.com/). Integration involved adding AI-generated individuals and relationships in OWL format, which were subsequently visualized and validated in Protégé.

Validation followed an expert-driven approach. A domain specialist (who was the developer of both the reference model and the original ontology individuals, with over 15 years of HTA experience) evaluated the GenAI outputs. Previously defined individuals were used as a reference and benchmark for assessing the completeness and semantic adequacy of the generated content.

Factorial design

A factorial experimental design was implemented to evaluate how multiple variables influence the ability of GenAI models to generate valid ontology individuals. Based on an analysis of the problem involving both ontologies and GenAI systems, a set of relevant factors was identified, along with their levels for factorial design and the corresponding initial hypotheses, as summarized in Table 1. This design provides a structured framework to analyze how variables interact and jointly affect output quality.

Three use cases were established to represent realistic scenarios in health economic modeling, each simulating different stages of the evidence synthesis process. All were formulated using the PICO framework (43), which is commonly used in evidence-based medicine to define the problem (P), intervention (I), comparator (C), and outcomes (O).

In the first case (UC1), designed to address exploratory or early-stage modeling where specific evidence is scarce, GenAI received only the PICO-formulated research question. This case tests the model's ability to act as a preliminary conceptualization tool with minimal context.

In the second (UC2), the PICO question was accompanied by previously generated conceptual model individuals for PBD, representing a scenario where an initial disease structure is already established but requires systematic population of parameters.

Table 1. Factorial matrix of factors and levels aligned with HTA modeling challenges

Factor	Levels (Id)	Supporting hypotheses/HTA relevance
Use case	Only PICO (UC1) PICO+Model (UC2) PICO+Report (UC3)	UC1 evaluates the feasibility of early-stage conceptualization with scarce evidence. UC2 simulates the systematic population of established disease structures, whereas UC3 assesses the automation of data extraction from extensive technical reports.
Ontology serialization	XML (F1) JSON-LD (F2) Turtle (F3)	Different serializations may influence the technical reliability and interoperability of the generated evidence model.
Prompting style	Complete upfront (P1) Iterative (P2)	Iterative refinement is expected to enhance the methodological rigor and transparency of the expert-guided generation process.
Information source	With sources (S1) Without sources (S2)	Structured sources are critical to ensure evidence-based accuracy and to mitigate the risk of clinical hallucinations.
GenAI model	Perplexity (M1) ChatGPT-4.1 (M2)	Evaluating models' performance is essential to identify tools capable of meeting the stringent reliability requirements of regulated HTA domains.

https://doi.org/10.1017/S0266462326103754 Published online by Cambridge University Press

In the third (UC3), GenAI was given a complete published technical report (44), reflecting the real-world task of extracting structured data from extensive HTA documentation to automate the transition from already published evidence to computational implementation.

Other experimental factors included the prompting style (iterative vs. single prompt), the inclusion or exclusion of specific information sources, and the choice of GenAI model (Perplexity deep research ChatGPT-4.1). Using explicit information sources, such as references on biotinidase deficiency (45–53), allowed for comparison with free-text generation based solely on model knowledge. However, it was determined that the application of this factor made sense only in UC1: UC2 represented a scenario where it should be expected to have already collected proper references, and UC3 made use of a report with validated data and references.

A full-factorial combination of these factors would yield 72 experimental runs. However, serialization was added as an additional factor in a subsequent phase of the experimentation, when the version of Perplexity used in the previous experiments was not available anymore. Consequently, to ensure comparability, JSON-LD and Turtle were only tested with ChatGPT-4.1, thus reducing the number of experiments to 48. These figures were reduced in turn to 32 (16 runs for UC1, 8 for UC2, and 8 for UC3) due to the dependency among the information source factor and the use case.

Prompt design and interaction with GenAI

The interaction with GenAI models followed prompt engineering principles (35), aiming to maximize the quality, reliability, and reproducibility of outputs. Prompt design adopted a modular structure, where each module controlled a specific aspect of the interaction (context, tasks, or information) enhancing the generation of valid ontology individuals. For each experiment, the total GenAI response time (excluding human interaction) and the number of corrections performed were recorded.

The main modules were: (i) a "role module," assigning the model the role of a health economics expert specialized in HTA, ensuring consistent and domain-appropriate responses; (ii) a "context module," describing the ontology's structure and concepts, that is, clinical manifestations, interventions, parameters, utilities, and costs; (iii) a "PICO question module," guiding the model through the PICO framework; (iv) an "information sources module," used only in UC1, specifying whether external references or internal model knowledge should be applied; (v) a "task module," defining objectives and expected ontology individuals; and (vi) an "iteration module," determining whether the process followed a sequential or "one-shot" interaction. Iterative prompting, as noted in Velásquez Henao et al. (35), improves accuracy and reduces hallucinations by enabling stepwise corrections. The number of individuals generated per interaction depended on the prompting strategy. In the single-prompt (non-iterative) setting, the model generated multiple individuals in a single response. In contrast, under the iterative prompting approach, individuals were generated progressively across successive interactions, with each iteration producing a subset of the required individuals.

This modular architecture ensured flexibility and scalability across experimental conditions while maintaining methodological rigor. Figure 1 illustrates the general design, and full prompt details for each use case are provided in the supplementary material (Supplementary Material Appendix B).

An iterative approach was used in Use Cases 1 and 3 (UC1, UC3), where GenAI progressively generated all individuals needed to represent biotinidase deficiency. In Use Case 2 (UC2), which already included a conceptual model, a simplified iterative procedure was applied to complete missing parameters. Table 2 summarizes the iterations for each case.

During early tests, models struggled to interpret OWL syntax, which led us to develop a corrective prompt. By providing explicit examples based on existing type 1 diabetes mellitus individuals, the corrective version helped GenAI better understand the ontology structure and relationships. This enhancement reduced errors and enabled the semi-automated generation of consistent and semantically valid individuals.

Statistical analysis of results

The analysis evaluated each experimental scenario to assess the quality of individuals generated by the models through qualitative and quantitative approaches. Most variables (both those defining the problem and those describing outcomes) were categorical, taking values from limited discrete sets that facilitated interpretation and statistical processing.

Validation involved two stages. First, an expert reviewed each of the above 1300 generated individuals, classifying them into categories (risks, manifestations, and interventions) and identifying hallucinations. Attending to the observed problems in the generated individuals, and based on existing literature on hallucinations in large language models (54–56), the expert classified the identified hallucinations into four categories:

i. Structural hallucinations: generating entities or properties that do not exist in the reference ontology, violating schema constraints.
ii. Semantic hallucinations: combining valid ontology terms in logically inconsistent ways, for example, violating domain, range, or class disjunction constraints.
iii. Contextual hallucinations: producing outputs that are structurally and logically valid but fail to meet input specifications, such as interventions for populations or conditions not requested.
iv. Evidence-based (factuality) hallucinations: producing literals, numerical values, or citations that appear plausible but are unsupported or factually incorrect according to clinical evidence.

After the individual validation, a holistic experiment assessment was performed. The qualitative assessment of each experiment involved four variables rated on a five-point Likert scale (1 = strongly disagree, 5 = strongly agree): "Consistency," that is, internal coherence across individuals; "Relevance," that is, alignment with the research question; "Completeness," that is, coverage of all required concepts; and "Adequacy," that is, practical usability without major revisions.

Cramer's V was employed to assess associations between categorical variables, as it effectively captures nonlinear dependencies and accommodates features with multiple levels. This analysis was supplemented with contingency tables and heatmaps to further explore the direction of the identified relationships.

We assessed the internal consistency of the four Likert-rated quality dimensions using Cronbach's alpha. Given the low internal reliability, we did not aggregate them into a single composite score. Instead, we conducted a multivariate analysis of variance (MANOVA) to test for global effects of the experimental factors on the joint quality profile, using Pillai's trace as the primary statistic due to the small sample size and potential deviations from multivariate normality. To identify dimension-specific effects, we then estimated separate linear models for each outcome variable. All models were fitted using heteroscedasticity-robust (HC3) standard errors. Finally, we derived reduced, more parsimonious

https://doi.org/10.1017/S0266462326103754 Published online by Cambridge University Press

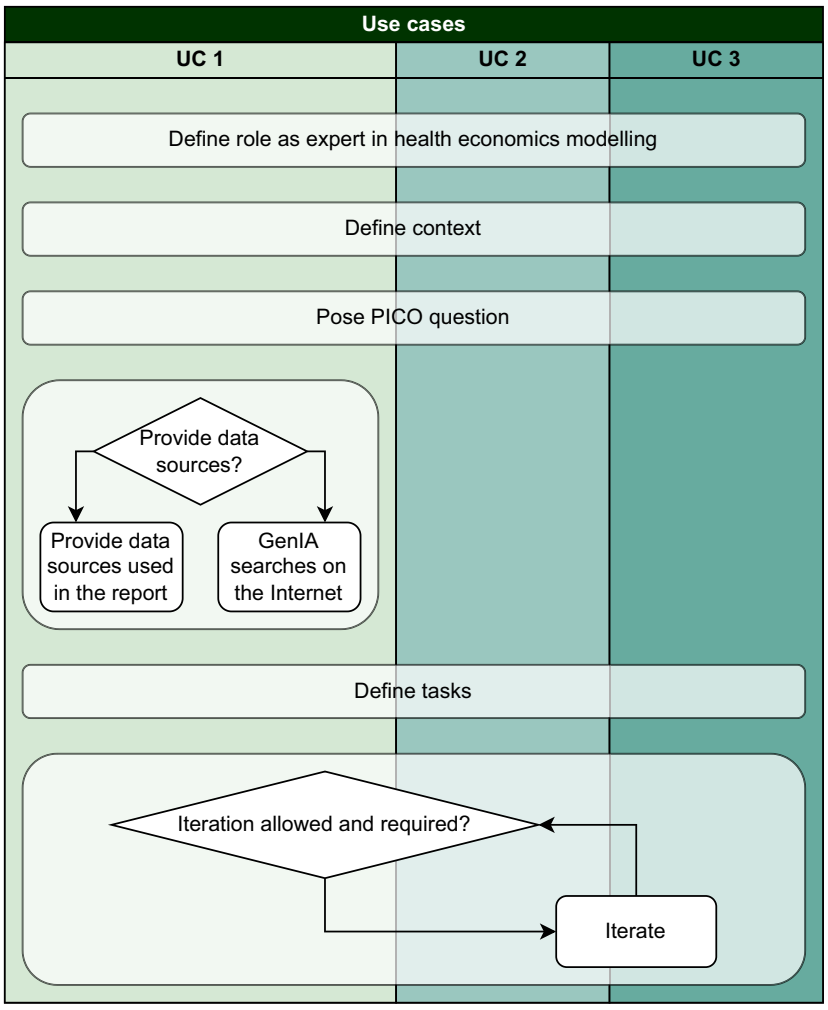


Figure 1. Modular structure of prompt design used in the experiments.

specifications by removing clearly non-informative predictors, balancing model fit (AIC/BIC), statistical stability, and consistency with the experimental design.

We finally explored the relationship between the experimental factors and the proportion of hallucinations. Because the dependent variable was a proportion, we fitted binomial logistic regression models with a logit link function, using the number of generated individuals as binomial weights. Given the evidence of overdispersion and consistently with previous analyses, we estimated heteroscedasticity-robust (HC3) standard errors.

A full list of variables used in the evaluation and statistical analyses is available in the supplementary material (Supplementary Material Appendix A).

Results and discussion

Two out of the 32 planned experiment runs had to be discarded because the output of the GenAI was impossible to parse and convert into a usable ontology format.

The preliminary analysis with Cramer's V (Figure 2) helped identifying a series of key findings:

i. The use case presents a strong association with the percentage of evidence hallucinations, number and coverage of the created individuals, but seems independent from other types of hallucinations, completeness, or adequacy.
ii. Providing sources is related to the coverage of concepts and, moderately, to relevance, but barely related to any other of the qualitative assessments.
iii. The prompting style shows a strong association with the number of created individuals and the coverage.
iv. The model used shows a moderate association with execution time and consistency.
v. Serialization is related to hallucinations (especially structural and evidence ones), the total qualitative score, and also to the number of corrections.

The analysis of the average qualitative assessment per experimental factor (Figure 3) was not conclusive due to the relatively small sample size but remarks a number of trends. It is observed that UC2 consistently outperforms the other cases in all categories. Regarding the model, ChatGPT shows higher average scores than Perplexity across all dimensions. Although serialization using JSON-LD scores higher in relevance, Turtle is the best performing serialization format in general. Providing sources scores better in every dimension but adequacy. The iterative prompting approach achieves a much higher score in completeness but performs similarly to the complete upfront approach in the rest of dimensions.

https://doi.org/10.1017/S0266462326103754 Published online by Cambridge University Press

Table 2. Prompts per iteration and use case

Iteration	Prompt description
Use case 1 and use case 3	
1	Generate main individuals to represent biotinidase deficiency and its stages based on the technical report.
2	Create individuals to represent birth prevalence for each stage of biotinidase deficiency using epidemiological data.
3	Generate individuals for acute clinical manifestations of the deficiency.
4	Generate individuals for chronic clinical manifestations of the deficiency.
5	Define probabilities of acute manifestations for the total deficiency condition.
6	Define probabilities of acute manifestations for the partial deficiency condition.
7	Generate individuals for utility and disutility values associated with each manifestation.
8	Specify temporal parameters such as onset times of acute and chronic manifestations.
9	Represent main interventions described in the report.
10	Represent treatments for each type of biotinidase deficiency.
11	Represent follow-up strategies for each type of biotinidase deficiency.
Use case 2	
1	Based on the provided reports and the ontology structure, generate progression probabilities for the following manifestations: • `PBD_ManifestationHearingProblems` • `PBD_ManifestationHypotonia` • `PBD_ManifestationMentalDelay` • `PBD_ManifestationSeizure` • `PBD_ManifestationSkinProblems` • `PBD_ManifestationVisionLoss` Ensure strict alignment with the conceptual model provided.
2	Generate utility values associated with each health state, based on the manifestations and the provided reports. Maintain the ontology's structure and consistency when incorporating the individuals.

The MANOVA reveals a significant multivariate effect of prompt style on the overall quality profile, whereas other factors show weaker or dimension-specific patterns. Subsequent univariate models clarify these effects: iterative prompting produces a substantial increase in completeness, model choice primarily affects consistency, and Turtle serialization is positively associated with adequacy (and, to a lesser extent, completeness). In contrast, relevance remains comparatively stable across experimental conditions. Overall, the results suggest that different experimental factors selectively influence distinct aspects of quality rather than producing a uniform improvement across all dimensions (see Supplementary Material Appendix A for details on the fitted models).

Regarding the proportion of hallucinations, the reduced model retains variables that exhibited substantial effect magnitudes in the full model, despite their initial lack of statistical significance. This selection criterion accounts for the constraints of the sample size and ensures that practically relevant factors are not overlooked. Once the model was refined, Turtle serialization was associated with a substantial decrease in the likelihood of hallucinations ($OR = 0.15$, p-value $< .001$), corresponding to an approximate 85 percent reduction in the odds relative to XML. No other experimental factor showed statistically significant effects, although a marginal positive tendency was observed for UC3.

An emergent finding was that hallucinations were more frequent in structurally complex ontology components, whereas well-defined classes were generated more reliably. For instance, GenAI typically struggles with defining parameters, such as costs, that involve second-order uncertainty. Such instances require classification as subclasses of both `Cost` and `SecondOrderParameter`. Furthermore, characterizing the uncertainty itself necessitates additional parameters to accurately model the underlying probability distribution. This pattern suggests that GenAI performance may reflect differences in ontological complexity, highlighting a potential complementary role in ontology evaluation.

The practical implementation of the experiments faced several limitations, mainly technical. The Perplexity API lacked features for controlled experimental setups, could not process files, and required manual insertion of ontology and source content due to token limits. Additional credit constraints further limited testing.

During early experiments, both models showed limitations in handling ontology files provided directly in OWL or RDF formats. This was identified through follow-up queries asking whether the ontology content had been successfully processed, to which the models indicated that they could not access or interpret the uploaded ontology files in their native format. As a result, ontology content was converted into plain text representations. Although this approach partially improved performance, the issue did not fully disappear, suggesting that additional factors such as input size, token limitations, or structural complexity may also have influenced the models' ability to process the information. Providing sample individuals and simplified representations ultimately proved to be the most effective workaround in our experimental setup.

Additionally, previous studies (57) have reported a decrease in the reliability of LLM-generated OWL outputs when multiple individuals are modeled simultaneously within a single prompt. In our experimental design, we did not explicitly evaluate this factor, as our focus was on other dimensions such as prompting strategy, serialization format, and information sources. Therefore, we did not observe clear evidence of this limitation under our specific setup, although the number of individuals generated per interaction was indirectly controlled through the iterative prompting strategy. This aspect represents an interesting direction for future work, particularly in relation to prompt scalability and output consistency.

We also observed occasional incomplete responses in which the model summarized groups of individuals and deferred the generation of remaining elements (e.g., "…include the remaining items here"), suggesting possible limitations related to response length or generation constraints. Although we did not explicitly analyze the effect of input or output size, these observations may be consistent with previous findings on performance degradation in more verbose formats such as XML. This aspect warrants further investigation in future work.

Manual expert validation of above 1,300 individuals was labor-intensive and potentially biased. Although alternative approaches, such as automatic ontology comparison (e.g., OntoSim (https://gitlab.inria.fr/moex/ontosim)) and GenAI-based validation, were explored, they proved impractical or unreliable for the requirements of this study.

https://doi.org/10.1017/S0266462326103754 Published online by Cambridge University Press

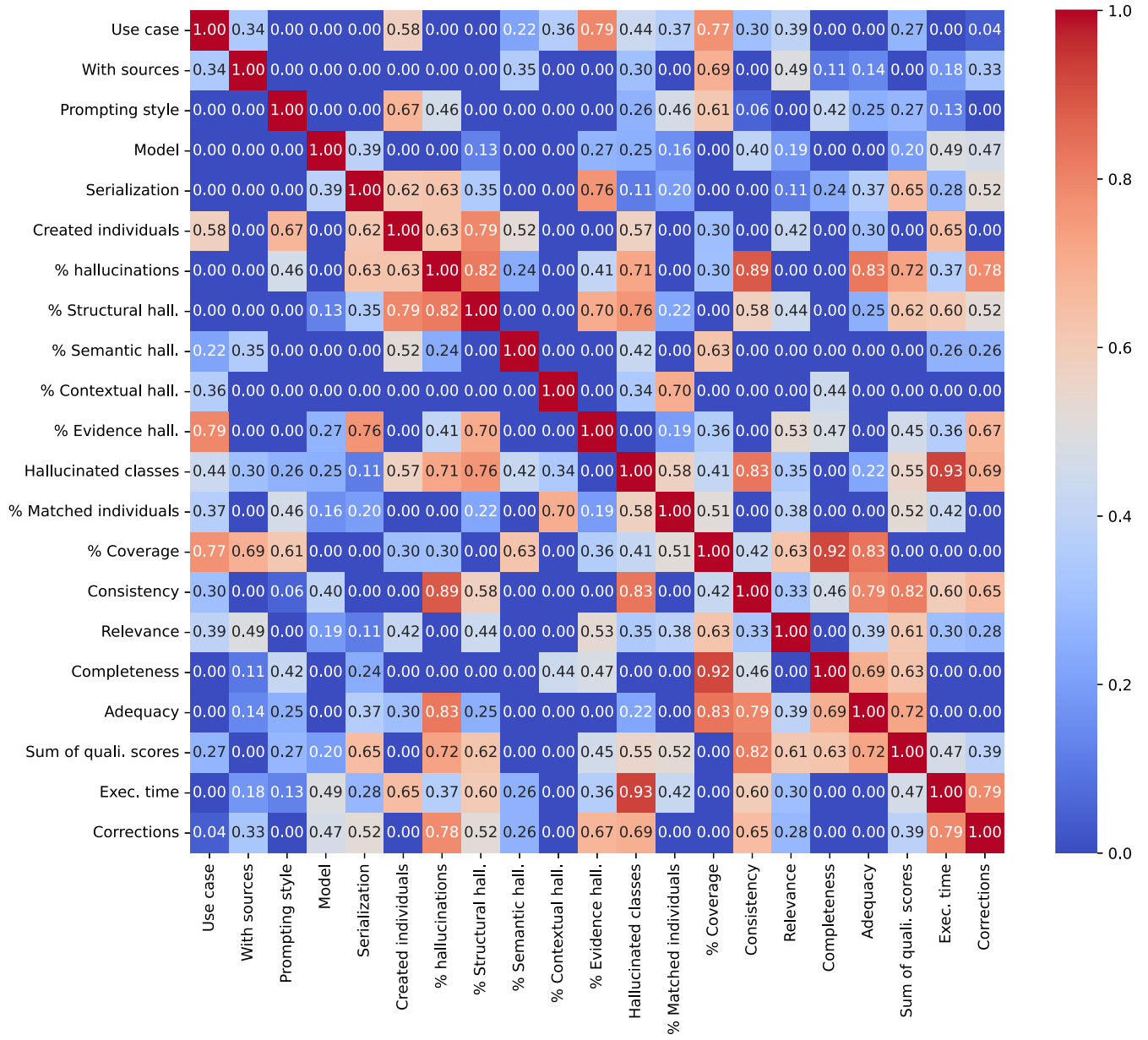


Figure 2. Cramer's V heatmap for the evaluated variables.

Conclusions and further work

This study evaluated the feasibility, reliability, and methodological implications of GenAI models for ontology population within HTA workflows. The results demonstrate that GenAI models can effectively generate structured ontology individuals, but their performance varies across distinct quality dimensions and is strongly influenced by experimental design choices. Rather than uniformly improving overall quality, factors such as prompting strategy and serialization format selectively affected completeness, adequacy, consistency, and the likelihood of hallucinations. In particular, iterative prompting significantly improved completeness, supporting its use as a controlled refinement mechanism during ontology generation. Additionally, Turtle serialization was associated with a substantially lower likelihood of hallucinations, suggesting that formal representation choices play a critical role in ensuring structural reliability and semantic validity (58;59;60).

For the HTA community, this approach offers a scalable solution for early-stage or exploratory modeling, where evidence is often dispersed and resource-intensive to synthesize manually. By providing a semi-automated pipeline to extract and structure clinical and economic knowledge, agencies and industry can reduce the time-to-insight without compromising methodological rigor.

However, the results also highlight important limitations. Both evaluated GenAI systems showed reduced reliability when generating individuals associated with more complex or abstract ontology structures, indicating that human oversight remains essential, especially for high-stakes or structurally complex modeling tasks. Furthermore, technical constraints related to model interfaces, input handling, and validation workflows currently limit full automation. These findings reinforce the importance of positioning GenAI as an assistive tool within expert-driven HTA processes rather than a fully autonomous solution.

https://doi.org/10.1017/S0266462326103754 Published online by Cambridge University Press

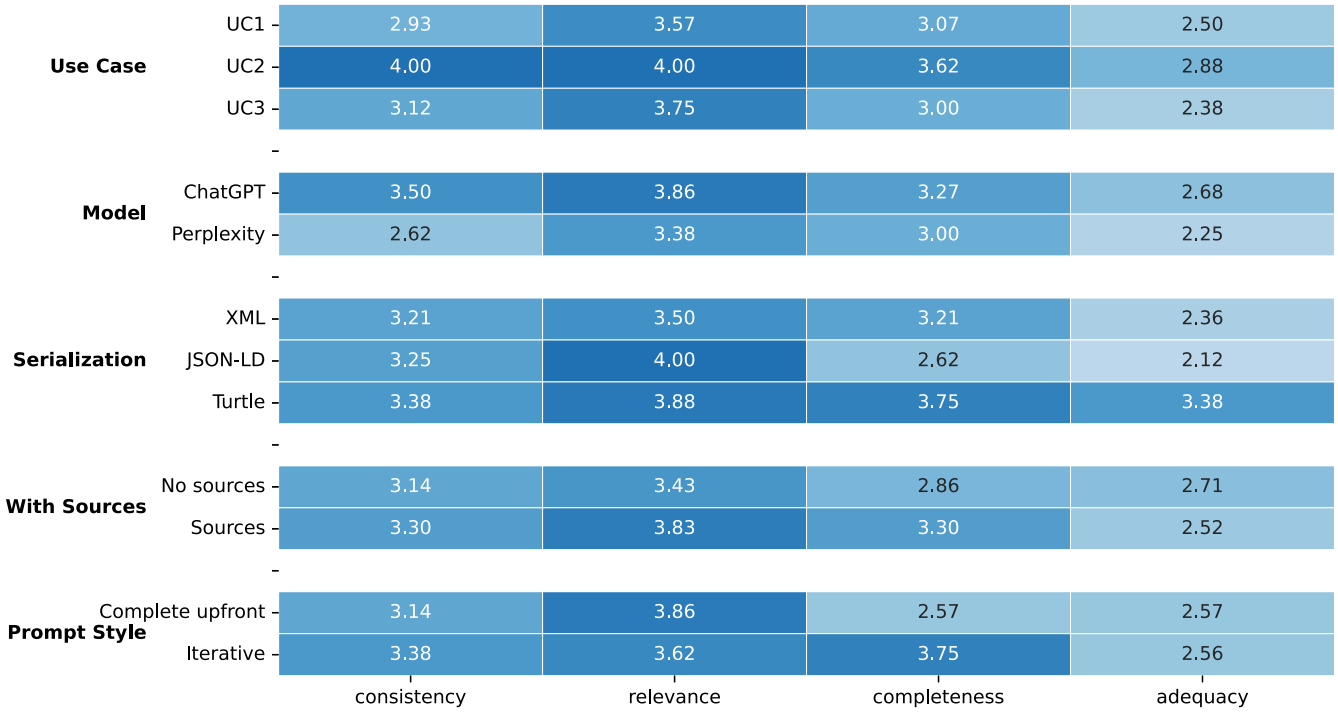


Figure 3. Heatmap of average qualitative assessment per experimental factor.

Future research aims to evolve this prototype into a production tool by broadening the validation scope. This evolution seeks to replace individual review with a panel of HTA experts to standardize the evaluative criteria and explore the use of AI-generated vignettes to support expert elicitation. We also propose mitigating the manual workload through automated tools like OntoSim and LLM-based alignment (e.g., *Agent-om* (61)), which would streamline the workflow and improve interoperability. Integrating comprehensive metrics as per Encord (62) will ensure the generative performance aligns with high-level quality standards.

Ultimately, integrating GenAI with formal semantic structures like OSDi provides HTA practitioners with a robust framework to manage the complexity of modern healthcare evidence, ensuring that automated outputs remain transparent, reproducible, and aligned with domain-specific standards. Lastly, evaluating the computational models automatically generated from these ontologies (a project already in progress (https://github.com/JaDES-ULL/JaDES-HTA)) will provide a complete evidence-to-model lifecycle of the proposed approach.

Supplementary materials. The supplementary material for this article can be found at http://doi.org/10.1017/S0266462326103754.

Data availability. All the data produced during the experiments described in this document, together with the details on the validation process, are publicly available at https://github.com/ontologies-ULL/OSDi-GenAI/.

Acknowledgements. The authors thank the researchers from the Evaluation and Planning Service of the Canary Islands Health Service (www.sescs.es) for their contribution to the definition of the PICO question and their comments during the validation process.

Author contribution.

- E.G.G. Conceptualization, Methodology, Writing – Original Draft.
- I.C.R. Conceptualization, Methodology, Validation, Formal Analysis, Writing – Original Draft.
- J.A.D.H. Software, Validation, Investigation, Data Curation, Writing – Original Draft.

Funding statement. No funding was received for conducting this study.

Competing interests. The authors have no relevant financial or nonfinancial interests to disclose.

References

1. **Briggs AH**, **Claxton K**, **Sculpher MJ**. *Decision modelling for health economic evaluation*. New York: Oxford University Press; 2006.
2. **Prieto-González D**, **Castilla-Rodríguez I**, **González-González EJ**, **de la Luz Couce-Pico M**. Automated generation of discrete event simulation models for the economic assessment of interventions for rare diseases using the RaDiOS ontology. *Int J Artif Intell Tools*. 2023;**32** (1):2350005. https://doi.org/10.1142/S0218213023500057.
3. **Kuziemsky CE**, **Lau F**. A four stage approach for ontology-based health information system design. *Artif Intell Med*. 2010;**50** (3):133-148. https://doi.org/10.1016/j.artmed.2010.04.012.
4. **Zhang Y**, **Gou L**, **shu Zhou T**, et al. An ontology-based approach to patient follow-up assessment for continuous and personalized chronic disease management. *J Biomed Inform*. 2017;**72**:45-59. https://doi.org/10.1016/j.jbi.2017.06.021.
5. **Alavi M**, **Leidner D**, **Mousavi R**. A knowledge management perspective of generative artificial intelligence. *J Assoc Inf Syst*. 2023;**25** (1):286-295. https://doi.org/10.17705/1jais.00849.
6. **Lee P**, **Bubeck S**, **Petro J**. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med*. 2023;**388** (13):1230-1236. https://doi.org/10.1056/NEJMsr2214184.
7. **Liu P**, **Yuan W**, **Fu J**, et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput Surv*. 2023;**55** (9):1-35. https://doi.org/10.1145/3560815.
8. **Bhuyan SS**, **Solanki V**, **Malik N**, et al. Generative artificial intelligence use in healthcare. *J Med Internet Res*. 2025;**13**:e52073.
9. **Biswas S**. Role of chat GPT in public health. *Ann Biomed Eng*. 2023;**51**:868. https://doi.org/10.1007/s10439-023-03172-7.

10. **Ouédraogo Z**, **Tapsoba LS**, **Sabane A**, et al. Text-to-OWL: Automated ontology construction for tuberculosis treatment recommendation using generative AI. In: Koné T, Sere A, Kouamé KF, editors. *Towards new e-infrastructure and e-Services for Developing Countries*. Cham: Springer Nature Switzerland; 2025, pp. 281-294.
11. **Brank J**, **Grobelnik M**, **Mladenic D**. A survey of ontology evaluation techniques. Proceedings of the 8th International Multi-Conference Information Society. 2005;p. 166-169.
12. **Castilla Rodríguez I**, **González E.** OSDi Repository. Available from: https://github.com/ontologies-ULL/OSDi.
13. **Moulaei K**, **Yadegari A**, **Baharestani M**, et al. Generative artificial intelligence in healthcare: A scoping review on benefits, challenges and applications. *Int J Med Inform*. 2024;**188**:105474. https://doi.org/10.1016/j.ijmedinf.2024.105474.
14. **Xu R**, **Wang Z**. Generative artificial intelligence in healthcare from the perspective of digital media: Applications, opportunities and challenges. *Heliyon*. 2024;**10** (12):e32364. https://doi.org/10.1016/j.heliyon.2024.e32364.
15. **Rouzrokh P**, **Alkhaldi S**, **Mohammadi R**. A current review of generative AI in medicine. *Front Med*. 2025;**12**:12185825. https://doi.org/10.1007/s12178-025-09961-y.
16. **Aggarwal T**, **Salatino A**, **Osborne F**, **Motta E**. Large language models for scholarly ontology generation. *Front Artif Intell*. 2026;**9**:1517918. https://doi.org/10.3389/frai.2025.1517918.
17. **Kollapally NM**, **Kotalawala SD**, **Dang HN**. Ontology enrichment using a large language model. *Artif Intell Med*. 2025;**144**:104818. https://doi.org/10.1016/j.artmed.2025.104818.
18. **Taboada M**, **Rivas G**, **Martinez J**. Ontology matching with large language models and prioritised depth-first search. *Inf Fusion*. 2025;**101**:103254. https://doi.org/10.1016/j.inffus.2025.103254.
19. **Reason T**, **Rawlinson W**, **Langham J**, et al. Artificial intelligence to automate health economic modelling: A case study to evaluate the potential application of large language models. *PharmacoEconomics Open*. 2024;**8**:191-203. https://doi.org/10.1007/s41669-024-00477-8.
20. **Chustecki M**. Benefits and risks of AI in health care: Narrative review. *Interact J Med Res*. 2024;**13**:e53616. https://doi.org/10.2196/53616.
21. **Panteli D**, **Adib K**, **Buttigieg S**, et al. Artificial intelligence in public health: Promises, challenges, and an agenda for policy makers and public health institutions. *Lancet Public Health*. 2025;**10** (5):e428-e432. https://doi.org/10.1016/S2468-2667(25)00036-2.
22. **Fleurence RL**, **Bian J**, **Wang X**, et al. Generative artificial intelligence for health technology assessment: Opportunities, challenges, and policy considerations: An ISPOR working group report. *Value Health*. 2025;**28** (2):175-183. https://doi.org/10.1016/j.jval.2024.10.3846.
23. **Reason T**, **Klijn S**, **Rawlinson W**, et al. Using generative artificial intelligence in health economics and outcomes research: A primer on techniques and breakthroughs. *PharmacoEconomics - Open*. 2025;**9**:501-517. https://doi.org/10.1007/s41669-025-00580-4.
24. **Qureshi R**, **Shaughnessy D**, **Gill KAR**, et al. Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? [note]. *Syst Rev*. 2023;**12** (1). https://doi.org/10.1186/s13643-023-02243-z.
25. **Reason T**, **Benbow E**, **Langham J**, et al. Artificial intelligence to automate network meta-analyses: Four case studies to evaluate the potential application of large language models. *PharmacoEconomics - Open*. 2024;**8** (2):205-220. https://doi.org/10.1007/s41669-024-00476-9.
26. **Li J**, **Deng Y**, **Sun Q**, et al. Benchmarking large language models in evidence-based medicine. *IEEE J Biomed Health Inform*. 2025;**29** (9):6143-6156. https://doi.org/10.1109/JBHI.2024.3483816.
27. **Gartlehner G**, **Kahwati L**, **Hilscher R**, et al. Data extraction for evidence synthesis using a large language model: A proof-of-concept study. *Res Synth Methods*. 2024;**15** (4):576-589. https://doi.org/10.1002/jrsm.1710.
28. **Schopow N**, **Osterhoff G**, **Baur D**. Applications of the natural language processing tool ChatGPT in clinical practice: Comparative study and augmented systematic review. *JMIR Med Inform*. 2023;**11**:e48933. https://doi.org/10.2196/48933.
29. **Szabó G**, **Pinsent A**, **Slim M**, et al. MSR179 automated extraction of cost-effectiveness models data from health technology assessment submissions using large-language models (LLMS): Does the prompting approach matter? *Value Health*. 2024;**27** (12):S473. https://doi.org/10.1016/j.jval.2024.10.2413.
30. **Meskó B**. Prompt engineering as an important emerging skill for medical education and practice. *J Med Internet Res*. 2023;**25** (1):e50638. https://doi.org/10.2196/50638.
31. **Ng JY**. Prompt engineering for generative artificial intelligence chatbots in health research: A practical guide for traditional, complementary, and integrative medicine researchers. *Integr Med Res*. 2025;**14** (4):101222. https://doi.org/10.1016/j.imr.2025.101222.
32. **Wang MH**, **Jiang X**, **Zeng P**, et al. Balancing accuracy and user satisfaction: The role of prompt engineering in AI-driven healthcare solutions. *Front Artif Intell*. 2025;8. https://doi.org/10.3389/frai.2025.1517918.
33. **Ambalavanan R**, **Snead RS**, **Marczika J**, et al. Ontologies as the semantic bridge between artificial intelligence and healthcare. *Front Digital Health*. 2025;7. https://doi.org/10.3389/fdgth.2025.1668385.
34. **Howell MD**. Generative artificial intelligence, patient safety and medical error: The foundation model paradigm. *BMJ Qual Saf*. 2024;**33** (11):748-756. https://doi.org/10.1136/bmjqs-2024-020032.
35. **Velásquez Henao JD**, **Franco Cardona CJ**, **Cadavid L**. Prompt engineering: A methodology for optimizing interactions with AI-language models in the field of engineering. *DYNA*. 2023;**90** (230):9-17. https://doi.org/10.15446/dyna.v90n230.111700.
36. **Suárez-Figueroa MC**. NeOn methodology for building ontology networks: Specification, scheduling and reuse [Doctoral dissertation]. Universidad Politécnica de Madrid. Madrid, Spain; 2010.
37. **Wolf B**. Clinical issues and frequent questions about biotinidase deficiency. *Mol Genet Metab*. 2010;**100** (1):6-13. https://doi.org/10.1016/j.ymgme.2010.01.003.
38. **Vallejo-Torres L**, **Castilla I**, **Couce ML**, et al. Cost-effectiveness analysis of a National Newborn Screening Program for Biotinidase deficiency. *Pediatrics*. 2015;**136** (2):e424-e432. https://doi.org/10.1542/peds.2014-3399.
39. **Eisenbarth GS**. Type 1 diabetes mellitus. A chronic autoimmune disease. *N Engl J Med*. 1986;**314** (21):1360-1368. https://doi.org/10.1056/NEJM198605223142106.
40. **Castilla-Rodríguez I**, **Arnay R**, **González-Cava JM**, et al. Towards an adaptive decision-support system for type I diabetes treatment based on simulation and machine learning. In: Bruzzone AG, Frascio M, Longo F, Novak V, editors. *Proceedings of the 8th international workshop on innovative simulation for healthcare (IWISH 2019)*; 2019, pp. 15-21.
41. **Shvets O**, **Murtazin K**, **Piho G**, **Meeter M**. Experiment with ChatGPT: Methodology of first simulation. *Front Educ*. 2025;10. https://doi.org/10.3389/feduc.2025.1624516.
42. **Musen MA**. The protégé project: A look back and a look forward. *AI Matters*. 2015;**1** (4):4-12. https://doi.org/10.1145/2757001.2757003.
43. **Santos CMdC**, **Pimenta CAdM**, **Nobre MRC**. Estrategia PICO Para la construcción de la pregunta de investigación y la búsqueda de evidencias. *Rev Lat Am Enfermagem*. 2007;**15**:508-511. https://doi.org/10.1590/S0104-11692007000300023.
44. **Vallejo Torres L**, **Castilla-Rodríguez I**, **Cuéllar Pompa L**, et al. Cost-effectiveness analysis of neonatal screening for biotinidase deficiency. Servicio de Evaluación del Servicio Canario de la Salud (SESCS); 2013. Technical report (in Spanish). Available from: https://sescs.es/coste-efectividad-del-cribado-neonatal-de-deficiencia-de-biotinidasa/.
45. **Salbert BA**, **Astruc J**, **Wolf B**. Ophthalmologic findings in biotinidase deficiency. *Ophthalmologica*. 1993;**206** (4):177-181. https://doi.org/10.1159/000310387.
46. **Joshi S**, **al Essa MA**, **Archibald A**, **Ozand PT**. Biotinidase deficiency: A treatable genetic disorder in the Saudi population. *East Mediterr Health J*. 1999;**5** (6):1213-1217.
47. **Wolf B**, **Spencer R**, **Gleason T**. Hearing loss is a common feature of symptomatic children with profound biotinidase deficiency. *J Pediatr*. 2002;**140** (2):242-246. https://doi.org/10.1067/mpd.2002.121938.
48. **Möslinger D**, **Mühl A**, **Suormala T**, **Baumgartner R**, **Stöckler-Ipsiroglu S**. Molecular characterisation and neuropsychological outcome of 21 patients with profound biotinidase deficiency detected by newborn screening and family studies. *Eur J Pediatr*. 2003;**162**:S46-S49. https://doi.org/10.1007/s00431-003-1351-3.

https://doi.org/10.1017/S0266462326103754 Published online by Cambridge University Press

49. **Weber P**, **Scholl S**, **Baumgartner ER**. Outcome in patients with profound biotinidase deficiency: Relevance of newborn screening. *Dev Med Child Neurol*. 2004;**46** (7):481-484. https://doi.org/10.1111/j.1469-8749.2004.tb00509.x.
50. **Genç GA**, **Sivri-Kalkanoglu HS**, **Dursun A**, et al. Audiologic findings in children with biotinidase deficiency in Turkey. *Int J Pediatr Otorhinolaryngol*. 2007;**71** (2):333-339. https://doi.org/10.1016/j.ijporl.2006.11.001.
51. **Ye J**, **Wang T**, **Han L**, et al. Diagnosis, treatment, follow-up and gene mutation analysis in four Chinese children with biotinidase deficiency. *J Inherit Metab Dis*. 2009;**32** (S1):295-302. https://doi.org/10.1007/s10545-009-1238-1.
52. **Couce M**, **Pérez-Cerdá C**, **Silva M**, et al. Hallazgos clínicos y genéticos en pacientes con deficiencia de biotinidasa detectados en el cribado neonatal o selectivo de sordera o de enfermedades metabólicas hereditarias. *Med Clin*. 2011;**137** (11):500-503. https://doi.org/10.1016/j.medcli.2011.01.018.
53. **Grünewald S**, **Champion MP**, **Leonard JV**, **Schaper J**, **Morris AAM**. Biotinidase deficiency: A treatable leukoencephalopathy. *Neuropediatrics*. 2004;**35** (4):211-216. https://doi.org/10.1055/s-2004-821080.
54. **Ji Z**, **Lee N**, **Frieske R**, et al. Survey of hallucination in natural language generation. *ACM Comput Surv*. 2023;**55** (12):1-38. https://doi.org/10.1145/3571730.
55. **Huang L**, **Yu W**, **Ma W**, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans Inf Syst*. 2025;**43** (2):1-55. https://doi.org/10.1145/3703155.
56. **Alkaissi H**, **McFarlane SI**. Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*. 2023;**15** (2):e35179. https://doi.org/10.7759/cureus.35179.
57. **Soares FM**, **Saraiva AM**, **Pires LF**, et al. Exploring a large language model for transforming taxonomic data into OWL: Lessons learned and implications for ontology development. *Data Intelligence*. 2025;**7** (2):265-302. https://doi.org/10.3724/2096-7004.di.2025.0020.
58. **Cao M**, **Wang Q**, **Zhang X**, et al. Large language models' performances regarding common patient questions about osteoarthritis: A comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and perplexity. *J Sport Health Sci*. 2025;14. https://doi.org/10.1016/j.jshs.2024.101016.
59. **Bashah A**, **Salem A**, **Al-waqeerah A**, et al. Evaluation of deepseek, gemini, ChatGPT-4o, and perplexity in responding to salivary gland cancer. *BMC Oral Health*. 2025;**25** (1). https://doi.org/10.1186/s12903-025-06726-4.
60. **Niel O**, **Dookhun D**, **Caliment A**. Performance evaluation of large language models in pediatric nephrology clinical decision support: A comprehensive assessment. *Pediatr Nephrol*. 2025;**40** (10):3211-3218. https://doi.org/10.1007/s00467-025-06819-w.
61. **Qiang Z**, **Wang W**, **Taylor K**. Agent-OM: Leveraging LLM agents for ontology matching. *Proc VLDB Endow*. 2024;**18** (3):516-529. https://doi.org/10.14778/3712221.3712222.
62. **Yu L**, **Alégroth E**, **Chatzipetrou P**, **Gorschek T**. Measuring the quality of generative AI systems: Mapping metrics to quality characteristics — snowballing literature review [review]. *Inf Softw Technol*. 2025;186. https://doi.org/10.1016/j.infsof.2025.107802.

https://doi.org/10.1017/S0266462326103754 Published online by Cambridge University Press