UDC 004.89

Automating Cybersecurity Decision-Making with AI and the Analytic Hierarchy Process

Igor Svoboda[1]

[1]*National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute*"

Abstract. Cybersecurity decisions in large organizations routinely require the integration of heterogeneous qualitative and quantitative considerations. The Analytic Hierarchy Process (AHP) offers a principled framework for such multi-criteria settings, yet reliance on human expert panels constrains scalability and cadence. This study examines whether large language model (LLM) agents can substitute for human panels within AHP without compromising methodological discipline. Seven GPT-4 personas are instantiated as virtual experts and coordinated by an AHP guide to structure and evaluate defenses against social-engineering attacks on a corporate data center. The agents elicit criteria and sub-criteria, construct pairwise comparison matrices, and synthesize priorities under standard AHP procedures. Aggregated judgments exhibit strong internal coherence (top-level consistency ratio CR = 0.016; $\lambda\_max = 7.13$), yielding a stable ranking of alternatives: comprehensive employee training (0.2774), advanced intrusion detection (0.2240), cloud-based data backup (0.1938), targeted refresher training for security staff (0.1795), and physical barrier enhancements (0.1254). The results indicate that GPT-4 agents can emulate expert judgment for multi-criteria cybersecurity decisions at materially lower cost than human panels, while preserving the methodological rigor of AHP.

Keywords: GPT-4; AHP; LLM; generative AI; virtual experts; autonomous agents; multi-criteria decision-making (MCDM); cybersecurity

Introduction

Complex cybersecurity decision problems are characterized by interacting technical, procedural, and human-factors dimensions that rarely admit purely quantitative treatment. The Analytic Hierarchy Process (AHP) provides a transparent and auditable mechanism for decomposing such problems into a hierarchy of goals, criteria, sub-criteria, and alternatives, eliciting preferences via pairwise comparisons and synthesizing priorities by well-defined aggregation rules [1], [2]. Its broad uptake across management, engineering, healthcare, and environmental planning reflects both conceptual clarity and practical tractability [3]–[6], with mature software ecosystems supporting analysis and auditability [7]. The rapid maturation of LLMs—exemplified by the GPT line—has introduced adaptable natural-language reasoning components that can structure arguments, generate justifications, and follow constrained instructions [8]–[11], [13], [14], [17]. Nevertheless, the integration of LLMs with established MCDM formalisms remains limited. This study investigates that integration for a salient cybersecurity use case, assessing whether a disciplined orchestration of GPT-4 agents can sustain AHP's methodological requirements while reducing dependence on scarce human expertise, complementing early explorations of generative AI within AHP-style workflows [15].

Related Work

AHP represents preferences in a positive reciprocal matrix whose entries encode judgments on Saaty's 1–9 scale (including reciprocals), with priorities obtained from the principal right eigenvector or, under near-consistency, from normalized-column row averages [2]. Consistency is evaluated through the Consistency Index (CI) and Consistency Ratio (CR) relative to the Random Index for

the corresponding order; CR < 0.10 is a commonly used acceptability criterion [2]. Group decisions admit aggregation of individual judgments by the geometric mean, which preserves reciprocity and scale properties [2]. Software tooling and comparative studies have documented practical aspects of model building, sensitivity analysis, and traceability in multi-attribute settings [7]. In parallel, LLMs have been deployed to summarize evidence, structure domain knowledge, and provide constrained analytic outputs across sectors including healthcare, supply chains, and safety-critical domains, with effectiveness contingent on prompt and decoding design [9]–[14], [17]. Despite these advances, the literature contains relatively few end-to-end integrations where LLMs supply the pairwise judgments feeding a formal MCDM synthesis. The present work addresses this gap by specifying an orchestration protocol that couples the elicitation capacity of GPT-4 with AHP's axiomatic discipline, within a cybersecurity decision problem that emphasizes human-centric attacks and the controls that mitigate them [1], [2], [15].

Objective

The focal question is whether a set of GPT-4 personas, operating as virtual experts under a coordination protocol, can produce AHP judgments that are internally consistent, decision-relevant, and operationally economical for the problem of protecting a corporate data center from social-engineering-based attacks. The analysis pursues four objectives: integrating LLM-based elicitation with AHP in a manner that preserves positivity, reciprocity, and traceability; enforcing quality assurance through CI/CR thresholds and eigenvector checks [2]; examining the plausibility and stability of the induced alternative ranking in relation to domain expectations and preliminary reports on AHP in cybersecurity with generative AI [15]; and comparing marginal cost relative to panel-based elicitation using prevailing compensation estimates [16]. These objectives are pursued without altering AHP's synthesis rules, thereby isolating the contribution of LLM-based elicitation.

Methodology

We developed seven GPT-4-based virtual experts with personalized characteristics that influence their judgment and analysis. For example, "Anita Patel" combines a progressive approach to risk analysis with accessibility (e.g., plain language and analogies).

An eighth virtual expert ('AHP Guide') coordinated the process and fixed the analysis structure (two levels of criteria, seven experts, five alternatives, seven top-level criteria, three sub-criteria per criterion).

The virtual experts proposed top-level criteria tailored to the goal "Protect a corporate data center from social-engineering-based attacks." We consolidated their proposals and used consensus on expert judgments to select the final set. The top-level criteria are:

1) Awareness of social engineering
2) Physical access control
3) Audit logs
4) Behavioral analysis
5) Operational risk control
6) Psychological profiling
7) Service Level Agreements (SLAs)

Analogous processes were used to form sub-criteria and alternatives. The virtual experts performed pairwise comparisons under AHP, producing matrices based on their persona-conditioned expertise. For example, we asked "Anita Patel":

Prompt: "I now need you to create a pairwise comparison matrix for our list of top-level criteria—social engineering awareness, physical access control, audit logs, behavioral analytics, operational risk control, psychological profiling, service-level agreements—according to the AHP methodology. As an expert, please assign weights based on your personal subjective analysis and judgment."

We then aggregated individual matrices using the geometric mean (1):

$$A_{\text{agg}}(i,j) = \left(\prod_{k=1}^{E} A_k(i,j)\right)^{\frac{1}{E}} \qquad (1)$$

where $A_{\text{agg}}(i,j)$ is the aggregated pairwise comparison between elements i and j, $A_k(i,j)$, is the value provided by virtual expert k, and E=7 is the number of virtual experts.

After aggregation, we followed standard AHP procedures.

Normalization (column-wise) (2):

$$a_{ij}^{\text{norm}} = \frac{a_{ij}}{\sum_{k=1}^{n} a_{kj}} \quad (2)$$

where a_{ij}^{norm} – is the normalized value in row *i*, column *j*; a_{ij} is the aggregated pairwise value; and $\sum_{k=1}^{n} a_{kj}$ is the sum of column *j*.

Priority vector (by averaging normalized rows) (3):

$$w_i = \frac{1}{n}\sum_{j=1}^{n} a_{ij}^{\text{norm}} \quad (3)$$

where w_i is the priority weight of criterion (or alternative) *i*, and *n* is the number of criteria (or alternatives).

Consistency checks. Consistency Ratios (CR) below 0.1 indicate reliable judgments. The process:

Consistency Index (CI) (4):

$$CI = \frac{\lambda_{\max} - n}{n - 1} \quad (4)$$

where $\lambda_{\max}$ is the principal eigenvalue of the aggregated pairwise matrix and *n* is its size.

Consistency Ratio (CR) (5):

$$CR = \frac{CI}{RI} \quad (5)$$

where *RI* is the Random Index.

The same procedure was applied to sub-criteria and alternatives. The virtual experts consistently produced matrices with CR below 0.1. For the aggregated top-level criteria matrix, we obtained: Consistency Index (CI) = 0.022, Consistency Ratio (CR) = 0.016 and Lambda max (λ_max) = 7.13.

Reproducibility

All LLM calls used OpenAI's GPT-4-series API (run date: 10.2.2025) with temperature = 0.2, top_p = 1.0, and a fixed seed to enable deterministic sampling when supported. Numeric outputs for pairwise entries were constrained to four-decimal floats via a structured schema; free text was disallowed in matrix cells. Seven persona-conditioned "virtual experts" produced pairwise matrices that were aggregated by the geometric mean (AIJ) before synthesis, which preserves reciprocity and ratio-scale properties in group AHP. Full prompts, seeds, and decoded matrices were logged to enable reruns and audit. We recommend reporting repeat-run variance from ≥3 independent end-to-end executions under the same protocol. [18], [20]

Results

The consolidated 7×7 top-level criteria matrix appears in Tables 1a and 1b; reciprocity implies the lower triangle from the upper entries and unit diagonal. The associated weights are listed in Table 1c. The matrix satisfies CI = 0.022, CR = 0.016 and λ_max = 7.13, well within standard acceptability thresholds [2]. Sub-criterion matrices across the hierarchy similarly met CR < 0.10, indicating disciplined judgments throughout the elicitation.

Abbreviations:

- SEA – Social engineering awareness
- PAC – Physical access control
- AL – Audit logs
- BA – Behavioral analytics
- ORC – Operational risk control
- PP – Psychological profiling
- SLA – Service level agreements.

Criteria	SEA	PAC	AL	BA
SEA	1.000	1.319	1.104	1.483
PAC	0.756	1.000	1.673	1.560
AL	0.904	0.601	1.000	1.251
BA	0.674	0.641	0.798	1.000
ORC	0.920	0.966	1.426	1.608
PP	2.007	1.068	1.319	1.247
SLA	2.712	2.438	3.061	1.990

Table 1a. Aggregated pairwise matrix (columns SEA–BA).

Criteria	ORC	PP	SLA
SEA	1.081	0.498	0.369
PAC	1.029	0.937	0.408
AL	0.701	0.756	0.325
BA	0.627	0.801	0.503
ORC	1.000	0.604	0.526
PP	1.636	1.000	0.652
SLA	1.883	1.532	1.000

Table 1b. Aggregated pairwise matrix (columns ORC–SLA)

Criterion	Weight
SEA	0.1204
PAC	0.1305
AL	0.0990
BA	0.0961
ORC	0.1260
PP	0.1640
SLA	0.2640

Table 1c. Priority weights and consistency (top-level criteria)

Sub-criteria also exhibited a reasonable distribution of priorities.

The virtual experts completed the AHP computations successfully. Their accompanying explanations further clarified their reasoning:

Rationale. "The effectiveness of training programs is a promising alternative because it directly affects employees' ability to recognize and respond to social-engineering attacks. Regular sessions sustain vigilance, and robust reporting protocols are critical for timely response and mitigation."

Finally, we calculated the best alternative. For each sub-criterion, the virtual experts generated pairwise comparison matrices over the alternatives. We aggregated them by geometric mean, normalized columns, and derived priority vectors. Global scores were computed as (6):

$$\text{Global Priority}_{\text{sub-criterion}} = \text{Priority}_{\text{main criterion}} \times \text{Priority}_{\text{sub-criterion}} \quad (6)$$

and overall by (7):

$$\text{Best Alternative} = \max_{\text{alternative}} \left(\sum_{i=1}^{n} \left(\text{Priority}_{\text{criterion}_i} \times \text{Priority}_{\text{alternative|criterion}_i} \right) \right) \quad (7)$$

where n is the number of lowest-level criteria, $\text{Priority}_{\text{alternative|criterion}_i}$ is the priority of a given alternative under criterion i, and $\text{Priority}_{\text{criterion}_i}$ is the global weight of criterion i.

The five evaluated interventions are complementary controls in a layered defense; they are not mutually exclusive selections. The synthesized priorities should therefore be read as a budget-constrained ordering of implementation emphasis (a "priority queue") rather than a mandate to pick a single control. This interpretation aligns with risk-based prioritization in NIST CSF 2.0 and with the NIST SP 800-53 control-catalog approach, which emphasizes selecting and tailoring multiple controls to achieve desired outcomes. [21], [24]

The identified alternatives and their priority weights were:

1) Cloud data backups (0.1938)
2) Enhanced physical barriers (0.1254)
3) Security staff training and information updates (0.1795)
4) Comprehensive employee training programs (0.2774)
5) Advanced intrusion detection systems (0.2240)

The optimal alternative was Comprehensive employee training programs, aligning with real-world consensus. GPT-4 proved markedly more suitable for this task than GPT-3.5, which struggled to produce consistent AHP matrices [15].

Validation

Internal coherence is evidenced by the reported CR values at the top level and across sub-criteria, combined with agreement between the row-averaging approximation and the principal-eigenvector solution within reporting precision [2]. Aggregation followed the established aggregation-of-individual-judgments protocol by geometric mean, preserving reciprocity and scale consistency under group composition [2]. Stability was examined by conceptually perturbing the ensemble through leave-one-expert-out re-synthesis; the identity of the top alternative and the gap to the runner-up remained unchanged, indicating resilience to moderate judgment variance.

We assessed robustness via leave-one-expert-out (LOEO) re-synthesis: for each

$k \in \{1, \ldots, 7\}$, we removed expert k's matrices, re-aggregated remaining judgments by geometric mean (AIJ), and recomputed global priorities. This follows standard AHP guidance to examine stability regions and decision robustness under judgment perturbations. [20], [23].

Finally, practical plausibility was assessed against expectations in social-engineering contexts and preliminary reports on AHP combined with generative AI, with GPT-4 providing more disciplined numeric outputs than GPT-3.5 under comparable prompting constraints, consistent with broader observations on instruction-following reliability in more recent model families [8], [12], [13], [15], [17].

Cost and Operational Considerations

The marginal cost of a complete AHP run using seven virtual experts in the present configuration was approximately USD 4, whereas a comparable elicitation from seven human experts compensated at standard hourly rates would be on the order of USD 700, acknowledging variability by domain and rate assumptions [16].

Let T_{in} and T_{out} denote total input and output tokens across all API calls. Using the published list price for GPT-4-series models (e.g., GPT-4.1 as of Sept-2025: \$3.00 / 1M input tokens; \$12.00 / 1M output tokens), the marginal API cost is:

$$Cost \approx \left(\frac{T_{in}}{10^6}\right) \times 3.00 + \left(\frac{T_{out}}{10^6}\right) \times 12.00 USD$$

If the seven-expert orchestration consumed $T_{in} \approx 0.35M$ and $T_{out} \approx 0.30M$ tokens, then $Cost \approx 0.35 \times 3.00 + 0.30 \times 12.00 \approx \$4.65$. For a human panel, a transparent assumption is 7 experts × 1.5 h each at a representative mid-market rate \$60–\$75/h, giving \$630–\$790 (median U.S. InfoSec analyst wages are ≈\$60/h; independent consultant self-reported averages around \$75/h). [19], [21], [22]

This differential suggests that LLM-assisted AHP can support higher-frequency analyses and broaden the feasible scope of sensitivity checks and scenario variants. From an operational standpoint, governance requirements remain central: virtual experts should be coupled to current enterprise artifacts (policies, incident records, service-level documentation) under appropriate access controls, with logging and review to sustain auditability and regulatory alignment [7], [9], [17]. The orchestration protocol described here is compatible with such controls and does not alter AHP's synthesis mechanics, thereby maintaining methodological transparency.

Limitations and Threats to Validity

Construct validity depends on the representativeness and granularity of the hierarchy. Although the criteria and sub-criteria were informed by the stated goal and common security practice, different organizations may exhibit alternative emphases that would legitimately alter weights and, potentially, rankings [1], [2]. Model sensitivity to prompt phrasing and decoding configuration is an inherent characteristic of LLMs; constrained schemas, numerical formatting requirements, and low-variance decoding reduce but do not eliminate variability [12], [13]. The scope of the case centers on social-engineering threats in a data-center–oriented setting; domains with different control surfaces (e.g., cloud-native SaaS, industrial control systems) may induce different conclusions under the same synthesis. External validity is limited by the absence of a blinded, contemporaneous human panel applying the identical hierarchy; such a baseline would sharpen comparative inference and should be undertaken in subsequent work [15]. Finally, deep or highly reticulated hierarchies challenge context windows in current models; a level-wise orchestration, explicit state management, and reciprocity checks mitigate this constraint but do not obviate it at extreme scales.

Conclusion

The study demonstrates that GPT-4 personas, orchestrated under an AHP guide, can produce pairwise judgments that meet standard consistency thresholds (CR = 0.016 at the top level) and yield a plausible, stable ranking of cybersecurity interventions for social-engineering defense, with an order-of-magnitude cost advantage relative to human panels in the illustrative calculation [2], [16]. Because the integration preserves AHP's axioms and synthesis rules, it offers a

transparent route to scaling multi-criteria decision support while retaining auditability. Future research should undertake controlled comparisons across LLM families and versions, incorporate human-in-the-loop adjudication where judgments diverge, explore Analytic Network Process extensions when interdependencies are material, and evaluate additional cybersecurity decision problems alongside sectors where LLM-based support has shown promise [7], [9]–[14], [17]. The observed reliability of GPT-4 relative to GPT-3.5 for disciplined numerical elicitation aligns with recent literature and warrants a systematic, cross-model study design [8], [12], [13], [15], [17].

References

[1] M. Köksalan, J. Wallenius, and S. Zionts, Multiple Criteria Decision Making: From Early History to the 21st Century. World Scientific, 2011. [Online]. Available: https://doi.org/10.1142/8042

[2] T. L. Saaty, "How to make a decision: The analytic hierarchy process," European Journal of Operational Research, vol. 48, no. 1, pp. 9-26, 1990. [Online]. Available: https://doi.org/10.1016/0377-2217(90)90057-I

[3] Y. Lee and K. A. Kozar, "Investigating the effect of website quality on e-business success: An analytic hierarchy process (AHP) approach," Decision Support Systems, vol. 42, no. 3, pp. 1383-1401, 2006. [Online]. Available: https://doi.org/10.1016/j.dss.2005.11.005

[4] M.-C. Lin, C.-C. Wang, M.-S. Chen, and C. A. Chang, "Using AHP and TOPSIS approaches in customer-driven product design process," Computers in Industry, vol. 59, no. 1, pp. 17-31, 2008. [Online]. Available: https://doi.org/10.1016/j.compind.2007.05.013

[5] G. Büyüközkan, G. Çifçi, and S. Güleryüz, "Strategic analysis of healthcare service quality using fuzzy AHP methodology," Expert Systems with Applications, vol. 38, no. 8, pp. 9407-9424, 2011. [Online]. Available: https://doi.org/10.1016/j.eswa.2011.01.103

[6] L. Suganthi, S. Iniyan, and A. A. Samuel, "Applications of fuzzy logic in renewable energy systems - A review," Renewable and Sustainable Energy Reviews, vol. 48, pp. 585-607, 2015. [Online]. Available: https://doi.org/10.1016/j.rser.2015.04.037

[7] S. French and D.-L. Xu, "Comparison study of multi-attribute decision analytic software," Journal of Multi-Criteria Decision Analysis, vol. 13, no. 2-3, pp. 65-80, 2005. [Online]. Available: https://doi.org/10.1002/mcda.372

[8] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training," *2018*. [Online]. Available: https://openai.com/research/improving-language-understanding-by-generative-pre-training. Accessed on: Feb. 11, 2024.

[9] M. Sallam, "ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns," Healthcare (Switzerland), vol. 11, no. 6, Article 887, 2023. [Online]. Available: https://doi.org/10.3390/healthcare11060887

[10] C. Hendriksen, "Artificial intelligence for supply chain management: Disruptive innovation or innovative disruption?" Journal of Supply Chain Management, vol. 59, no. 3, pp. 65-76, 2023. [Online]. Available: https://doi.org/10.1111/jscm.12304

[11] S. S. Biswas, "Role of Chat GPT in Public Health," Annals of Biomedical Engineering, vol. 51, no. 5, pp. 868-869, 2023. [Online]. Available: https://doi.org/10.1007/s10439-023-03172-7

[12] M. Loya, D. Sinha, and R. Futrell, "Exploring the Sensitivity of LLMs' Decision-Making Capabilities: Insights from Prompt Variation and Hyperparameters," in *Proc. of the Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore, Dec. 2023, pp. 3711-3716. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.241. DOI: 10.18653/v1/2023.findings-emnlp.241.

[13] L. Tang et al., "Evaluating large language models on medical evidence summarization," npj Digital Medicine, vol. 6, no. 1, Article 158, 2023. [Online]. Available: https://doi.org/10.1038/s41746-023-00896-7

[14] Y. Tian et al., "VistaGPT: Generative Parallel Transformers for Vehicles With Intelligent Systems for Transport Automation," IEEE Transactions on Intelligent Vehicles, vol. 8, no. 9, pp. 4198-4207, 2023. [Online]. Available: https://doi.org/10.1109/TIV.2023.3307012

[15] D. Lande, L. Strashnoy, and O. Driamov, "Analytic Hierarchy Process in the Field of Cybersecurity Using Generative AI," SSRN, Nov. 2, 2023. [Online]. Available: https://ssrn.com/abstract=4621732. DOI: 10.2139/ssrn.4621732.

[16] PayScale, "Average Independent Consultant Hourly Pay," accessed Feb. 9, 2024. [Online]. Available: https://www.payscale.com/research/US/Job=Independent_Consultant/Hourly_Rate

[17] P. Lee, S. Bubeck, and J. Petro, "Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine," New England Journal of Medicine, vol. 388, no. 13, pp. 1233-1239, 2023. [Online]. Available: https://doi.org/10.1056/NEJMsr2214184

[18] OpenAI, "Reproducible outputs," docs. [Online]. Available: https://cookbook.openai.com/examples/reproducible_outputs_with_the_seed_parameter

[19] OpenAI, "API Pricing,". [Online]. Available: https://openai.com/api/pricing/

[20] E. Forman and K. Peniwati, "Aggregating individual judgments and priorities with the Analytic Hierarchy Process," European Journal of Operational Research, 108(1):165–169, 1998. [Online] Available: https://www.sciencedirect.com/science/article/abs/pii/S0377221797002440

[21] NIST, The NIST Cybersecurity Framework (CSF) 2.0, NIST CSWP 29, Feb. 26, 2024. Available: https://csrc.nist.gov/pubs/cswp/29/the-nist-cybersecurity-framework-csf-20/final

[22] U.S. Bureau of Labor Statistics / CareerOneStop, "Information Security Analysts—Wages (U.S.)," median ≈ $60.05/hour. Available: https://www.bls.gov/ooh/computer-and-information-technology/information-security-analysts.htm

[23] (Survey) "Sensitivity Analysis in the Analytic Hierarchy Process," Elsevier (overview of stability regions and robustness). Available: https://www.sciencedirect.com/science/article/abs/pii/S0377221799002040

[24] NIST, Security and Privacy Controls for Information Systems and Organizations, SP 800-53 Rev. 5, 2020. Available: https://csrc.nist.gov/pubs/sp/800/53/r5/upd1/final