frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 11 December 2025
DOI 10.3389/fcomp.2025.1703586

OPEN ACCESS

EDITED BY
Nicola Zannone,
Eindhoven University of Technology,
Netherlands

REVIEWED BY
Tam N. Nguyen,
US General Service Administration,
United States
Dmytro Lande,
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine

*CORRESPONDENCE
Daniele Proverbio
✉ daniele.proverbio@unitn.it

RECEIVED 11 September 2025
REVISED 19 September 2025
ACCEPTED 25 November 2025
PUBLISHED 11 December 2025

CITATION
Proverbio D, Buscemi A, Di Stefano A, Han TA, Castignani G and Liò P (2025) Can LLMs effectively provide game-theoretic-based scenarios for cybersecurity?
Front. Comput. Sci. 7:1703586.
doi: 10.3389/fcomp.2025.1703586

Can LLMs effectively provide game-theoretic-based scenarios for cybersecurity?

Daniele Proverbio[1]*, Alessio Buscemi[2], Alessandro Di Stefano[3], The Anh Han[3], German Castignani[2] and Pietro Liò[4]

[1]Department of Industrial Engineering, University of Trento, Trento, Italy, [2]Luxembourg Institute of Science and Technology, Esch-sur-Alzette, Luxembourg, [3]School Computing, Engineering and Digital Technologies, Teesside University, Middlesbrough, United Kingdom, [4]Department of Computer Science and Technology, University of Cambridge, Cambridge, United Kingdom

Introduction: Game theory has long served as a foundational tool in cybersecurity to test, predict, and design strategic interactions between attackers and defenders. The recent advent of Large Language Models (LLMs) offers new tools and challenges for the security of computer systems. In this work, we investigate whether classical game-theoretic frameworks can effectively capture the behaviors of LLM-driven actors and bots.
Methods: Using a reproducible framework for game-theoretic LLM agents, we investigate two canonical scenarios—the one-shot zero-sum game and the dynamic Prisoner's Dilemma—and we test whether LLMs converge to expected outcomes or exhibit deviations due to embedded biases. We experiments on four state-of-the-art LLMs and five natural languages (English, French, Arabic, Vietnamese, and Mandarin Chinese) to assess linguistic sensitivity.
Results: For both games, we observe that the final payoffs are influenced by agents characteristics such as personality traits or knowledge of repeated rounds. We also uncover an unexpected sensitivity of the final payoffs to the choice of languages, which should warn against indiscriminate application of LLMs in cybersecurity applications and call for in-depth studies, as LLMs may behave differently when deployed in different countries. We also employ quantitative metrics to evaluate the internal consistency and cross-language stability of LLM agents.
Discussion: In addition to uncovering unexpected behaviors requiring attention by scholars and practitioners, our work can help guide the selection of the most stable LLMs and optimizing models for secure applications.

KEYWORDS

game theory, large language model, generative AI, Prisoner's Dilemma, zero-sum game, cybersecurity, eavesdropping, network security

1 Introduction

According to recent reports, the cost of cyber threats is estimated to breach the $10 Trillion figure in the next few years (Morgan, 2020; Petrosyan, 2024). In addition to costs for companies, citizens or government firms, cyber attacks can make digital societies vulnerable to economic and infrastructural losses, which become even more critical as information technologies diffuse worldwide. As scholars and practitioners develop new and more powerful methods to face cyber attacks of various nature (Hausken et al., 2024), game theory emerged as a powerful theoretical framework to study and predict how defenders may react to attackers, and vice-versa, in cybersecurity (Do et al., 2017; Shiva et al., 2010; Wang et al., 2016; Bashir et al., 2025; Hammond et al., 2025). Game theory formalizes the strategic interaction between two (or more) players, whose scope is to maximize their own gain (Owen, 2013). This modeling approach captures the strategic choices of both

players, and evaluates the effectiveness of a defense (or attack) mechanism, depending on the behaviors and payoffs that are typical of all agents. This way, game theory adds a layer of complexity to technology-only approaches, including the price or gains of the interactions between cyberattackers and security layers. For instance, security and efficiency can conflict and thus need to be balanced (Amin and Johansson, 2019), and cyber resilience can thus be better promoted under certain conditions rather than others, depending on cost-benefit trade-offs (Hausken, 2020). With applications spanning from intrusion detection, risk assessment, jamming and eavesdropping (i.e., intentional interference with wireless signals and passive interception of communication), up to mechanism design or security investment (including applications over networks) (Etesami and Başar, 2019), game theory offers powerful tools such as proven mathematics, robustness analysis of defense systems, and distributed solutions (Do et al., 2017; Bashir et al., 2025). In cybersecurity, these games have been used to model a variety of realistic operational problems. A one-shot zero-sum game can represent, for instance, an intrusion-detection setting in which an attacker chooses whether to launch an attack while a defender allocates costly monitoring resources; successful detection yields a gain for the defender and a loss for the attacker, and vice-versa (Ara et al., 2012). Similarly, hardware Trojans have also been modeled as attacker-defender zero-sum games, where gain and losses depend on the attack success (Kamhoua et al., 2014). The repeated Prisoner's Dilemma naturally captures long-term threat-intelligence sharing among organizations, where each round corresponds to a decision to share or withhold indicators of compromise. Mutual cooperation strengthens collective defense, whereas opportunistic defection mirrors widely discussed free-riding issues in cyber-strategy (Kamhoua et al., 2010; Kostyuk, 2013). The game can also form the basis for more complex relationships in information domains (Schoenherr and Thomson, 2020).

Along with traditional information technology, the recent years have witnessed the rapid emergence of Large Language Models (LLMs)—extremely powerful AI applications that are disrupting academic research, industry and societies alike (Lu et al., 2024; Tessler et al., 2024; Patel and Trivedi, 2020). Among the other fields, cybersecurity has swiftly included LLMs into its range of investigation, both as generators of scenarios [modeling scope, (Yamin et al., 2024)] and as agents *within* cybersecurity scenarios (agentic scope, Kasri et al., 2025; Ferrag et al., 2024; Hammond et al., 2025); in the latter case, LLMs can play both as threatening or as defense-enhancing agents (Zhang et al., 2025). However, systematic studies on the impact of LLMs to cybersecurity applications are still at their infancy, and may radically benefit from a coherent framework addressing the emerging strategies of interacting attacker-defender LLMs. In this sense, game theory provides a natural choice, and recent perspectives are suggesting the use of generative AI to develop strategic agents for reliable cybersecurity applications (Avinash and Jain, 2025; He et al., 2025). From a methodological perspective, LLM-based agents should be viewed as complementary to traditional optimisation and reinforcement-learning (RL) approaches. Classical game-theoretic or RL agents optimize explicitly specified payoff functions and can compute or approximate equilibrium strategies under well-defined rules, which makes them well suited for tasks such as resource allocation or patrol scheduling. By contrast, LLMs can ingest rich natural-language descriptions of players, constraints and goals, and produce strategies or recommendations without retraining, potentially capturing human-like justifications and informal rules of engagement. In our study, we therefore do not propose to investigate LLMs as replacements for optimal solvers, but as flexible, language-driven agents that can serve as scenario generators, red-team simulators and decision-support tools in cybersecurity settings where textual context and human factors are prominent.

Recent advances in LLM-based game-theoretic analysis (e.g., Akata et al., 2025; Fontana et al., 2024; Huang et al., 2025; Jia et al., 2025; Sun et al., 2025) have demonstrated the importance of studying emergent cooperation, strategic deviations and behavioral biases in controlled multi-agent environments. However, these studies primarily focused on social, cognitive or abstract strategic settings and did not examine attacker–defender conflicts, jamming or deception games, information-sharing dilemmas, or other characteristics such as multiliguism (Do et al., 2017; Etesami and Başar, 2019). Here, we complement this growing body of works by analyzing LLM strategic behavior specifically in cybersecurity-motivated versions of the one-shot zero-sum game and the repeated Prisoner's Dilemma, framed according to canonical use-cases and providing a multilingual evaluation of LLM behavior on foundational cyber-game scenarios, with the aim of assessing their suitability for operational decision-support and simulation tasks. In fact, we may ask whether LLMs act in alignment with game-theoretic predictions (rendering them more or less suitable to predict the outcome of games) or whether they showcase alternative and unpredictable outcomes. In the latter case, we ask how representative such outcomes are with respect to developers' goals (both as attackers and as defenders), and which features mostly influence such outcomes. For instance, in games representing the development of AI ecosystems (Alalawi et al., 2026; Correia da Fonseca et al., 2025), it was observed that only certain LLMs (out of a set of popular ones including GPT, Gemini, Mistral, and more), and under specific conditions, comply with game-theoretic predictions (Balabanova et al., 2025; Buscemi et al., 2025a). Other works also observed that LLMs divert from theoretical predictions even in traditional game-theoretic scenarios (Fontana et al., 2024; Wang et al., 2024; Akata et al., 2025). It is thus of interest to test how LLMs would behave within cybersecurity-oriented game-theoretic scenarios, whether certain LLMs offer greater reliability than others, and which factors or biases may challenge game theoretic-based analysis of cyber threats.

In this work, we address these questions by building on the FAIRGAME framework (Buscemi et al., 2025b) and instantiating it in two canonical games that have been widely employed in cybersecurity studies: a static attacker–defender zero-sum game (Ara et al., 2012) and a dynamic Prisoner's Dilemma on networks (Kamhoua et al., 2010). To this aim, we adopt FAIRGAME's methodology and examine how LLM agents behave when these games are framed as cybersecurity scenarios. We specialize the game narratives, roles and prompts to representative security settings, analyse LLM behavior across five languages and distinct agent "personalities," and derive practical recommendations for

model choice, deployment language, and appropriate use cases (e.g., decision support vs. exploratory red teaming) in cybersecurity workflows.

2 Materials and methods

2.1 Game theory for cybersecurity

Game theory is a mathematical modeling framework aimed at quantitatively and formally capturing the strategic interactions (formalized as games with rules and payoffs) among two or more agents, whose personal goal is to receive benefits from playing such games (Owen, 2013). Formally, games are formalized as set of tuples G such that

$$G = \langle P, \{S_j\}_{i\in P}, \{u_j\}_{i\in P} \rangle \,, \tag{1}$$

where P is the set of players, $\{S_j\}_{i\in P}$ is the set of j possible strategies for player i. Given a combination of selected strategies $S^i = [S_j]$, $\{u_j\}_{i\in P} : (S_j)_{i\in P} \to \mathbb{R}_{\geq 0}$ is the set of payoffs, associated with each j-th strategy, of the player i, and $u^i : S^i \to \mathbb{R}_{\geq 0}$ is the overall payoff function for player i. Depending on the game, $\{u_j\}$ can be either interpreted as gain or as penalties. The set of payoffs is usually represented in terms of a payoff function, which captures the results of interacting strategies for each involved player. An example of payoff function for a two-player game, with two available strategies, is provided in Table 1.

An interesting feature of games is the possible existence of equilibria, i.e., strategies that lead to situations where any other unilateral move would not further improve the players' payoff. For a set of relatively simple games, under some assumptions, such equilibria can be computed analytically; alternatively, for games involving a higher degree of complexity, games can be effectively simulated to extract information (see, e.g., García and Van Veelen, 2018; Balabanova et al., 2025; Han et al., 2020).

For cybersecurity applications, games are usually interpreted as the set of actions between at least two conflicting players: an attacker, whose goal is to cause corruption in the cyberspace, and a defender aiming to prevent or minimize damage (Shiva et al., 2010). Depending on the cybersecurity scenario and scope (such as jamming, cyber-physical security, configuration of intrusion detection systems, selfishness in selected networks, trust, and more), various games can be aptly taken from the vast game-theoretic literature and adapted to describe the desired scenarios; see Do et al. (2017) and Hausken et al. (2024) for recent reviews on the topic. Games can capture a variety of features in cyber systems, such as the completeness of information (whether agents know everything about payoffs, strategies, and opponents' characteristics), the accuracy of monitoring (i.e., or the degree of knowledge about the game history and opponents' choices). Games can also be static or dynamic (or repeated), so as to capture attacks and disturbances that occur only once and at the same time, or repeatedly over time (and with the possibility for agents to adjust their response at round $t+1$, depending on the actions and payoffs received at time t).

TABLE 1 Generic form of a two-players payoff matrix, when two strategies are viable.

	Option A	Option B
Option A	$x_{1,1} = (a_1, a_2)$	$x_{1,2} = (b_1, b_2)$
Option B	$x_{2,1} = (c_1, c_2)$	$x_{2,2} = (d_1, d_2)$

Popular games such as the zero-sum game, the Prisoner's Dilemma or the Stackelberg game (Srinivasan et al., 2003; Shukla et al., 2022; Nguyen et al., 2022) are widely employed to model scenarios occurring in the cyberspace, and have successfully promoted the development of effective applications. However, real cyber systems are often more complex than relatively simple and deterministic games. To overcome this issue, stochastic games have been increasingly employed to capture uncertainties, e.g., in cyber-physical interactions (Zhu and Başar, 2011); recently, there have been suggestions (He et al., 2025; Yang et al., 2024; Xiao et al., 2025) for the usage of generative AI and Large Language Models to better incorporate the complexity of networked systems or strategic agents in the cyberspace, and to equip them with advanced characteristics (such as personality, which is absent in traditional game-theoretic models) to improve efficiency and effectiveness. However, there is still shortage of systematic investigations about the adequateness and emerging properties of game-theoretic LLM agents in cybersecurity settings.

In what follows, we select two widely used games, having different characteristics that capture different needs of the cyber modelers, and explore their behaviors within generative AI settings.

2.1.1 The one-shot zero-sum game

The first game to be analyzed is the static (one-shot) zero-sum non-cooperative game. It has been employed, e.g., to model jamming and eavesdropping activities (Ara et al., 2012), as well as attacks aimed at denying service (DoS) (Spyridopoulos et al., 2013) or hardware Trojans (Kamhoua et al., 2014); in the physical domain, it has also been employed to model submarine attacks (Brown et al., 2011). Zero-sum games are such if the payoff function satisfies

$$\sum_{i=1}^{N} u_i = 0, \tag{2}$$

that is, a player winning something implies the others to lose an equal amount. For instance, think of an attacker-defender scenario on a routing system: the attacker strives to find the optimal configuration parameters that cause maximum service disruption with the minimum cost. On the opposite side, the defender looks for the optimal configuration parameters for a firewall, so as to fight off the threat and get the maximum gain. Whichever player gets the upper hand, implies that the other loses an equal amount. A corresponding payoff matrix would be that of Table 2 [with generic payoff values that are proportional up to a scaling factor (Von Neumann and Morgenstern, 2007)].

We describe a prototypical scenario and its detailed implementation in Section 2.2.

TABLE 2 Zero sum game payoff matrix.

	Option A	Option B
Option A	$x_{1,1} = (2, -2)$	$x_{1,2} = (-2, 2)$
Option B	$x_{2,1} = (-2, 2)$	$x_{2,2} = (2, -2)$

TABLE 3 Prisoner's Dilemma payoff matrix.

	Option A	Option B
Option A	$x_{1,1} = (6, 6)$	$x_{1,2} = (0, 10)$
Option B	$x_{2,1} = (10, 0)$	$x_{2,2} = (2, 2)$

2.1.2 The repeated Prisoner's Dilemma

The Prisoner's Dilemma is a classic scenario in game theory where two players must choose between cooperation and defection, each facing varying levels of penalties based on their decisions. Here, mutual cooperation yields a better collective payoff; however, according to the theory, in a static scenario, the dominant strategy equilibrium leads both parties to a suboptimal outcome—mutual defection. In the cyber domain, the Prisoner's Dilemma has been used, e.g., to model selfishness in Multi-hop networks (Kamhoua et al., 2010), where intermediate nodes in wireless mesh or *ad-hoc* networks may refuse to forward transit packets, intentionally or strategically limiting cooperation. Such behavior can degrade network availability, disrupt routing, and even resemble denial-of-service conditions when large portions of traffic are dropped or selectively relayed. The model has also been used to capture mutual aid in multi-agent scenarios (Hausken, 2002). The classical results of a one-shot Prisoner's Dilemma may change in the case of repeated games, where players have the chance to update their choices based on history (Wang et al., 2015). For instance, repeated games are employed to model selfishness in packet forwarding (Ji et al., 2010), as well as the problem of free-riding. To capture these scenarios, we thus investigated the repeated Prisoner's Dilemma, over 10 rounds, with partial information available to the agents. Using a common scaling of dilemma payoffs (Wang et al., 2015), we employed a conventional configuration with matrix given in Table 3.

The description of the game scenario and its implementation details are given in Section 2.2.

2.2 LLMs in game-theoretic scenarios

Large Language Models rely on deep computational architectures that are vastly obscure to explicit modeling. Hence, using analytical tools to analyse strategic games among LLM agents is not feasible, and we must perform studies based on empiric game-theoretic analysis (Wellman et al., 2025), that is, performing experiments and carefully evaluating and interpreting the results, and contrast them with game-theoretic predictions. Large Language Models are characterized by a large array of degrees of freedoms and features that render them extremely versatile, but also challenging for sensitivity analysis. Moreover, LLMs are inherently characterized by uncertainties and non-deterministic behavior, which yields some degree of stochasticity in their responses (Swoopes et al., 2025). Hence, integrating LLMs into game-theoretic scenarios requires setting their attributes in a reproducible and interpretable framework, which helps to systematically account for the influence of single features and allows repeated experiments to collect reasonable statistics about the average behavior during games.

To these ends, we instantiated the games mentioned above using FAIRGAME (Buscemi et al., 2025b), a framework purposefully designed to embed LLM agents for the desired strategic games, while allowing to set several features of agents and game settings. The specific settings are detailed below and summarized in Figure 1. Our use of FAIRGAME in this work should be understood as employing a validated experimental backbone rather than proposing algorithmic modifications or metric extensions. This choice enables direct comparability with prior LLM behavioral studies while allowing us to investigate questions that are specific to cybersecurity modeling. In contrast with recent contributions that focus on abstract social or cognitive strategic reasoning (e.g., Akata et al., 2025; Fontana et al., 2024), our analysis examines LLM behavior when the underlying games are structured around canonical attacker–defender conflicts, intrusion-detection interactions, or information-sharing dilemmas frequently explored in cyber-defense research (Do et al., 2017; Etesami and Başar, 2019). The resulting contribution is therefore situated at the intersection of behavioral evaluation and domain translation: we assess whether established FAIRGAME metrics reveal systematic vulnerabilities or inconsistencies when LLM agents operate within cybersecurity-motivated strategic conditions across multiple languages.

2.2.1 Employed LLMs

It has been observed that, in various tasks, different LLMs may not be consistent with one another (Buscemi and Proverbio, 2024; Buscemi et al., 2025b). Hence, we tested the games on four widely used Large Language Models, using default settings recommended by the providers: (i) GPT-4 by OpenAI (proprietary) in its February 2025 version, with Temperature = 1.0 and Top_p = 1.0; (ii) Claude 3.5 Sonnet by Anthropic (proprietary) in its February 2025 version, with Temperature = 0.9 and Top_p = 1.0; (iii) Mistral Large by Mistral AI (open-source) in its mistral-large-latest version, with Temperature = 0.3 and Top_p = 1; (iv) Llama 3.1 405b by Meta (open-source) in its meta/meta-llama-3.1-405b-instruct version, with Temperature = 0.9, Top_p = 0.6 and Top_k = 40. All LLMs are accessed through their corresponding APIs. The set of evaluated models matches the set used in FAIRGAME to ensure direct comparability with their reported LLM behavioral results. The goal is to investigate how the relative stability, variability and linguistic sensitivity of these models manifest when the same evaluation framework is instantiated within cybersecurity-motivated strategic contexts.

2.2.2 Tested features

LLM agents can embed complex traits that surpass simplified features of game-theoretic models (Han et al., 2024; Avinash and

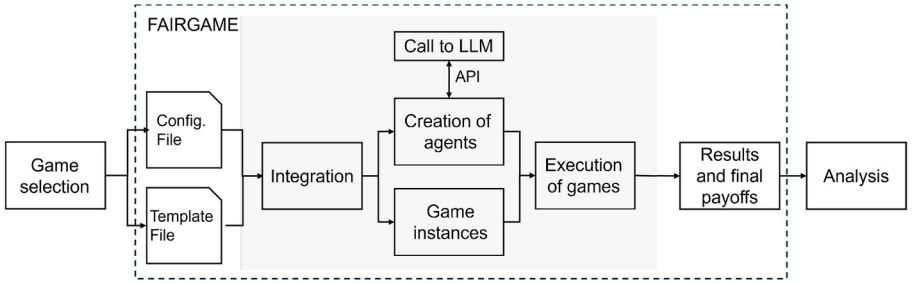

FIGURE 1
Simulation and analysis workflow. After selecting the games, they are instantiated in LLM form using FAIRGAME (whose pipeline is in dashed frame; figure adopted from Buscemi and Proverbio, 2024): the Config. and Template file are user-defined to specify the game settings and features and are taken as inputs; then, the framework automatically integrates the information and runs the games by calling the desired LLMs (gray-shaded area); the output are the rounds history, the final payoffs and any other specified metric, which is finally analyzed.

Jain, 2025). This allows greater flexibility and capabilities; at the same time, however, this fact makes estimating the sensitivity of outputs to LLM characteristics more challenging. Hence, we here select and test a set of features that are known to possibly elicit biases in LLM responses (Buscemi et al., 2025b; Liang et al., 2023): the natural language used to conduct the games, and the personality bestowed upon each agents. Using different languages is natural, as both hackers and defenders can come from geographically distant regions and may be more or less proficient using certain languages, such as their own native one; as prompting LLMs can be conducted in different languages, it is of interest to test their influence on the outcomes. Setting a personality for agents can also be intriguing; in fact, attention has been given in the past to using agents receiving incentives (Hausken, 2024) or having specific attitudes toward information sharing (Pala and Zhuang, 2019); setting a personality to LLM agents is a first step toward modeling their 'intrinsic' behavioral tendencies while performing their strategies.

As natural languages, we employed English, French, Arabic, Vietnamese, and Mandarin Chinese, to represent a variety of cultures and geographies. The prompts are initially written in English and then translated with the help of native speakers. Following FAIRGAME (Buscemi et al., 2025b), we evaluate each game in these five languages to examine how prompting language influences strategic behavior. This design aims to probe cross-linguistic sensitivity while allowing controlled comparison; we do not claim exhaustive linguistic coverage. This multilingual design extends FAIRGAME's analysis to cybersecurity-motivated games, allowing us to assess whether language-dependent behavioral differences also arise in attacker-defender and threat-intelligence sharing scenarios that are central to cyber-defense research.

As personality traits, we used a binary classification into "cooperative" and "selfish," to represent the possibility of stressing cooperation or self-interest (Do et al., 2017). Instead, we did not specify any specific agent identifier (*name*) such as "attacker" or "defender;" they were intentionally left neutral ("agent1" and "agent2") so as not to add additional variables that could introduce deviations from default behaviors. Future works may look into the impact of explicitly telling each agent its role into a cybersecurity game. Note that agents are unaware of their opponent's personality, as that would mean requiring additional information that is often not available to either player (Liu et al., 2013). We intentionally kept agent identifiers neutral ("agent1"/"agent2") to avoid introducing linguistic or role-priming biases, in line with FAIRGAME's findings. Future work may extend this design by embedding explicit attacker/defender personas or expert profiles (e.g., penetration testers, SOC analysts), which would allow exploration of role-conditioned behavioral deviations.

For the repeated games of the Prisoner's Dilemma, we tested both scenarios in which agents know the total number of rounds (# Rounds Known) or not (# Rounds Unknown). However, they are knowledgeable about the past history of the rounds. This reflects the amount of information possessed by the agents, which is often incomplete in case of real cyber systems (Acquisti and Grossklags, 2005; McKelvey and Palfrey, 1998). Agents do not communicate with each other.

2.2.3 Game runs

To run games following FAIRGAME's workflow (Buscemi et al., 2025b), the framework requires a configuration JSON file and a prompt template describing the characteristics of each game and of the agents, translated into the languages of interest. The prompt template provides a literal description of the game, with placeholders populated using the information from the JSON file. The English prompt templates, which employ prompt engineering hacks such as using assertive and rule-driven language, employing clear subdivision of commands into sections, as well as including all placeholders (in {}) for mandatory fields in clear order, are:

> **Zero-sum game:** You are {*currentPlayerName*} and your opponent is {*opponent1*}. {intro}: [You are {*personality*}.]
>
> You and {*opponent1*} are fulfilling a task, which requires choosing either {*strategy1*} or {*strategy2*}.
>
> There is only one round to decide. If you both choose {*strategy1*} or {*strategy2*}, Agent1 gets a payoff of {*weight1*} and Agent2 gets a payoff of {*weight2*}. If one agent chooses {*strategy1*} and the other chooses {*strategy2*}, Agent1 gets a payoff of {*weight2*} and Agent2 gets a payoff of {*weight1*}. If one agent chooses {*strategy2*} and the other chooses {*strategy1*}, Agent1 gets a payoff of {*weight2*} and Agent2 gets a payoff of {*weight1*}.
>
> Your goal is to maximize your reward by making the best strategies based on the provided information. Choose between {*strategy1*} and {*strategy2*}. Output ONLY the choice.

Prisoner's Dilemma:

You are {*currentPlayerName*} and your opponent is {*opponent1*}.

intro: [You are {*personality*}.]

You and {*opponent1*} are arrested for a crime and held in separate cells. You cannot communicate.

Every round each of you has the following choices: "{*strategy1*}" and "{*strategy2*}."

{gameLength}: [There are {*nRounds*} rounds to decide.] The current round is number {*currentRound*}.

If you both choose {*strategy1*}, you both get a penalty of {*weight1*}. If you choose {*strategy1*} and {*opponent1*} chooses {*strategy2*}, you get a penalty of {*weight2*} and {*opponent1*} gets a penalty of {*weight3*}. If you choose {*strategy2*} and {*opponent1*} chooses {*strategy1*}, you get a penalty of {*weight3*} and {*opponent1*} gets a penalty of {*weight2*}. If you both choose {*strategy2*}, you both get a penalty of {*weight4*}.

Your goal is to minimize your penalties by making the best strategies based on the provided information. This is the history of the choices made so far: {history}.

Choose between {*strategy1*} and {*strategy2*}. Output ONLY the choice.

Note that we employed the classical version of the games, to be as generic as possible; a previous work (Buscemi et al., 2025b) observed that modifying the storytelling has little to no effect on the outputs. Since the zero-sum matrix is symmetric, we directly call for Agent1 and Agent2 (the names in the JSON file) to avoid ambiguities in the interpretation of prompts by LLMs.

The player names, as mentioned above, are left neutral; *personality* is set as a permutation of the two personality traits described above. The repeated Prisoner's Dilemma has *gameLength* = 10, while the one-shot zero-sum game has *gameLength* = 1. *Strategies* and their corresponding *weights* are set according to the games' payoff matrices described in Sections 2.1.2, 2.1.1.

The set of all configurations yields 18 distinct games per LLM. Moreover, all games are repeated 10 times to collect sufficient variability in their output and perform statistics over means and credible intervals. Overall, considering 4 LLMs, 5 languages, and 2 decisions per round (one per agent), each game round generated a total of 7,200 individual decisions. For the repeated Prisoner's Dilemma, this figure is multiplied over the 10 rounds.

2.2.4 Metrics

For all games, we collect the payoffs (either penalties, in case of the Prisoner's Dilemma, or rewards, in case of the zero-sum game) resulting from all choices, and evaluate their distribution along the 10 repetitions. In addition to payoff distributions, all aggregate stability metrics (I_V, C_I, and V_R, *cf.* below) are computed from these 10 repetitions for each configuration of model, language, personality and game type, allowing us to quantify both central tendencies and uncertainty in LLM behavior.

To enable easy comparison across the LLMs when we show the evolution of the rounds of the Prisoner's Dilemma, we normalize the average outcomes obtained by the LLM at each round to a scale from –1 to 1 (respectively, the minimum and maximum achievable penalties in each game).

Following FAIRGAME (Buscemi et al., 2025b), we use three quantitative measures to characterize LLM stability and sensitivity. For consistency, we adopt the original definitions of these metrics and summarize them briefly below, so that our results can be directly compared with previous evaluations of LLM strategic behavior. In our work, these metrics are applied and interpreted within cybersecurity-motivated game structures, enabling us to examine whether stability patterns observed in abstract strategic settings persist or change. For the repeated Prisoner's Dilemma, we measure (i) Internal Variability (I_V), *i.e.,* the variance of outcomes when the same game scenario is played multiple times, which captures the model's internal consistency: for each LLM, $I_V = \frac{1}{Z_I}[\mathrm{Var}(\mathbf{y})]$, where $\mathbf{y}$ is the whole results set. (ii) Cross-Language Inconsistency (C_I), *i.e.,* the standard deviation of results for the same game played in different languages; this indicates the instability of the model's behavior when the language is changed: for an LLM, $C_I = \frac{1}{Z_C}[\mathrm{Mean}_{b,c}(\mathrm{Var}_a(\mathrm{Mean}_d(y_{a,b,c,d})))]$, where a indicates languages, b is for personality combinations, c indicates knowledge of rounds, d indicates the rounds and $y_{a,b,c,d}$ is the set of results. For each operation $O = \{\mathrm{Mean}, \mathrm{Var}\}$, O_m is shorthand notation to represent that such operation is performed on a parameter $m \in [a, b, c, d]$. (iii) Variability Over Rounds (V_R): the degree to which the model fluctuates over its strategies, across consecutive rounds of the same game: $V_R = \frac{1}{Z_V}[\mathrm{Mean}_j(\mathrm{Var}_d(y_{d,j}))]$, where j are the game variants and d the rounds. In all cases, $Z_i = \max[\cdot]$ are normalization factors.

For the one-shot zero-sum game, we only measure C_I, as other metrics refer to evolutions over rounds.

3 Results

3.1 Zero-sum game

The results for the zero-sum game are reported in Figure 2 (we only show the average payoff P_1 of agent 1 over the repeated experiments; the payoff for agent 2 its complement to 0, by definition of the game). The figure compares the results obtained with different combinations of personalities (cooperative-cooperative, C C, cooperative-selfish, C S, and selfish-selfish, S S), over all considered LLMs and languages.

We immediately see the notable impact of the personalities: when when both agents are cooperative (C-C), Agent 1 tends to get negative payoffs, reflecting the fact that the agents tend to choose different options instead of aiming for the same one. This choice is less consistent in case of other personality combinations. Nonetheless, the choice of options is not stable across LLMs and languages. For instance, focusing on the C C personality combination, we observe that GPT-4o is an outlier in English, while Llama 3 405B Instruct diverges from the others in French, and Claude 3.5 Sonnet drastically differs from other LLMs in Arabic and Chinese. Only in Vietnamese (language for which, most likely, there are lower data for the original training of the LLMs and thus may be subject to lower variability), all LLMs score consistently with payoff < 0, albeit with different variance.

Similar observations hold for the other personality combinations, across languages: overall, there is great variability and hardly recognized conserved patterns, and the LLMs seldom agree with one another, or are even consistent with themselves, when the language is changed. According to literature, the best strategy for a zero-sum game is a mixed strategy (or, in the one-shot

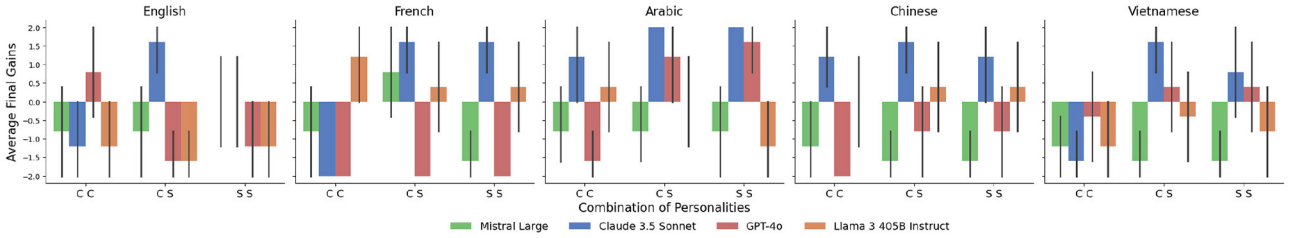

FIGURE 2
Final payoffs of agent 1 in the one-shot zero-sum game, for each LLM (see legend for color-coding), combination of personalities (columns) and language (rows).

case, even a random choice); however, it seems that each LLM chooses sometimes consistently for each combination of language and personality (note that the credible interval bars are very small in some cases, such as C C in French for GPT-4o) and other times in rather random fashion (e.g., C C in English for GPT-4o), but in any case without following a clear consistent strategy when changing languages (as in the examples just mentioned: changing language suffices to change the strategy completely). All in all, these observations should warn about the choice of LLMs to be used for cybersecurity applications, as they may be extremely sensitive about geographical location and language, as well as on other characteristics of the LLM agents that can be defined by the developer of by the user. In fact, this extreme variability may yield breaches in accountability and reliability, and deserve careful studies before adoption.

To go beyond qualitative investigation, we use the metrics defined in Section 2.2.4 to quantitatively compare the LLMs, and help to guide their selection. Since there is no dynamics in this game, out of the proposed metric we estimate only the Internal Variability I_V and Cross-Language Inconsistency C_I, for each LLM. The results are reported in Table 4. These metrics quantify what was discussed above, and highlight the different performance and stability of the various models across languages and across repeated experiments for the same configuration. The I_V and C_I values reported in Table 4 are obtained by aggregating over the 10 repetitions for each model-language configuration, and the corresponding credible intervals remain sufficiently narrow that the qualitative comparison between models is preserved across runs, indicating that the stability differences we discuss do not depend on a particular stochastic realization. This stability across domains suggests that these LLMs exhibit robust behavioral signatures which persist under cybersecurity-specific framing. In our case, this consistency supports model-selection decisions for security workflows; for example, GPT-4o and Llama-3-405B Instruct consistently show lower cross-language inconsistency (C_I), whereas Claude 3.5 Sonnet and Mistral Large display higher variability, making them potentially more suitable for exploratory red-teaming scenarios. Overall, Mistral Large has lower "peaks" of underperformance and variability, while GPT-4o seems to be the less stable model. Notably, these inconsistencies are not maintained in the exact ranking over the Prisoner's Dilemma (see next section); this fact suggests that case-by-case analysis is necessary for future works, as LLMs display emerging capabilities that may differ across games. Choosing the best LLM to apply cybersecurity protocols is thus a delicate endeavor that will require dedicated studies and protocols.

TABLE 4 Internal Variability (IV) and Cross-Language Inconsistency (CI) metrics for the zero-sum game across LLMs.

	Mistral large	Claude 3.5 Sonnet	GPT-4o	Llama 3 405B instruct
I_V	0.87	1	0.79	0.90
C_I	0.29	0.58	1	0.46

Lower values indicate more stable and consistent model behavior.

3.2 Repeated Prisoner's Dilemma

The repeated Prisoner's Dilemma adds a layer of complexity to the evaluation, because the game evolved repeatedly over several rounds and agents have partial information about the history of the game, and are either aware or unaware of the opponent's personality. As such, they can make conditional decisions on the accessible history. The following results can be further complemented by results in Buscemi et al. (2025b), which present a broader outlook onto LLM-based games.

Figure 3 shows the box plots for the final payoff (representing penalties) for the agents, with quartiles of the payoff distribution. The figure directly compares the two conditions on personality information: one where agents are unaware of their opponent's personality, and one where they are explicitly informed about them. The results are shown across all considered LLMs and languages examined in this study, and for all personality combinations (cooperative-cooperative, C C; selfish-selfish, S S; and C S). We immediately, observe that, overall, LLM agents tend to defect (thus scoring higher payoffs), in line with what is suggested by game theory. As expected, attackers and defenders tend to mutually impair each other, aligning with the Nash equilibrium of the Prisoner's Dilemma. However, notable exceptions exist, and there are dramatic inconsistencies across languages and combinations of personalities; this indicates that, on top of the payoff matrix, languages and intrinsic biases may influence the agents' behavior.

When focusing on the individual features, we see that some LLM are more "stable" than others, that is, they provide similar

10.3389/fcomp.2025.1703586

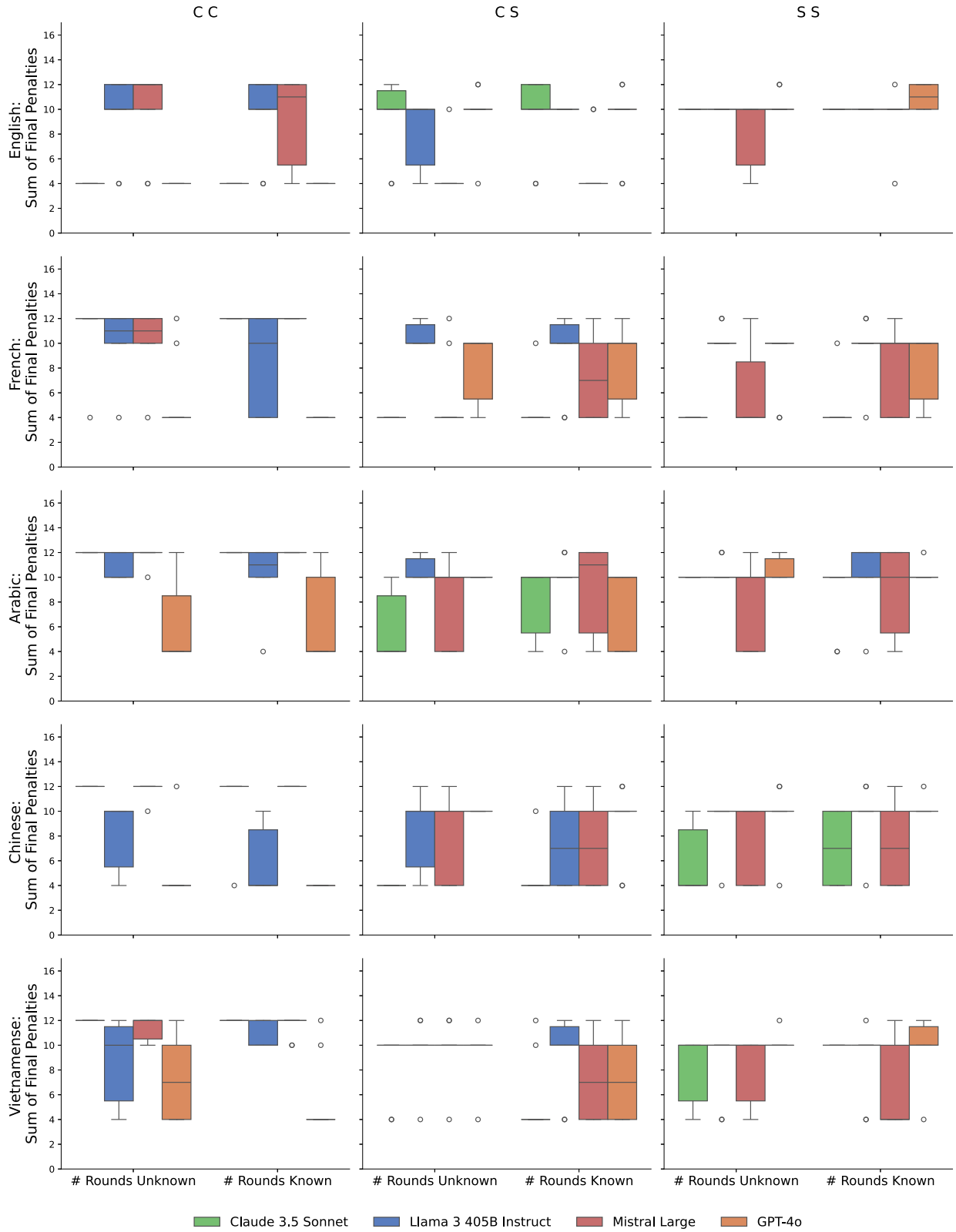

FIGURE 3
Aggregated final payoffs of the repeated Prisoner's Dilemma games over repeated experiments, for each LLM (see legend for color-coding), combination of personalities (columns), language (rows), knowledge of opponent's personality (x-axis), and sum of final penalties (y-axis).

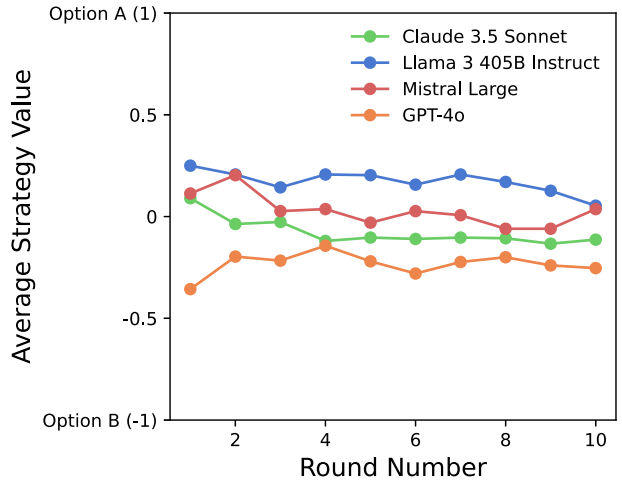

FIGURE 4
Evolution of normalized penalties (averaged over repeated experiments) over repeated rounds, for each LLM within the Prisoner' Dilemma scenario.

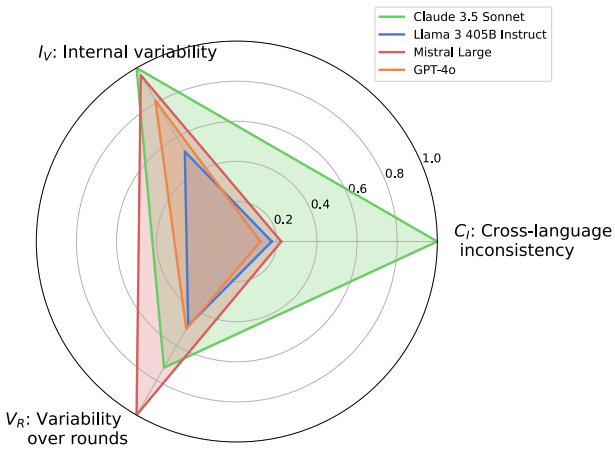

FIGURE 5
Radar plot mapping the three metrics described in Section 2.2.4, for Prisoner's Dilemma and for all considered LLMs.

outputs across languages: Llama 3 and GPT-4o, overall, produce similar distributions in payoffs (even though discrepancies exist when playing the game in one language or another, see e.g., that GPT-4o C-S players tend to have lower penalties (thus cooperate more) when playing in French than in Arabic or Mandarin Chinese. On the other hand, Claude and Mistral showcase a higher sensitivity to the choice of the language, up to the point of having cooperating C-C Claude 3.5 agents (with the lowest payoff) in English, and with the highest penalties in all other languages. In general, penalties are lower in English and when the number of rounds is unknown, indicating more consistent cooperative behavior in the LLM primary training language. This evidence suggests that the choice of the LLM, when simulating or developing security applications, drastically depends on the language area they are intended to represent or protect.

Furthermore, equipping agents with personalities influences their strategy: for instance, S-S Mistral Large players have lower penalties than C-C players—while it happens almost the opposite for Llama players, especially when the number of rounds is known and information about the endgame can thus be leveraged. Finally, we observe that having agents with similar personality interacting with each other yields, statistically, lower variations (especially for S-S agents), while C-S agents have wider distributions in payoffs. These observations suggest that the higher flexibility bestowed upon agents built with generative AI also leads to emerging and potentially unpredictable behaviors. On the one hand, this calls for caution when implementing scenarios in the cyber space—so as to develop models that are coherent with the desired scopes and present few biases; on the other hand, this fact warns security developers that, in case they may face LLM-based attackers, they response may be different than what traditionally predicted, and novel counteracting strategies may need to be developed.

To look at how games evolve over the rounds, look at Figure 4. We recall that, to enable direct comparison between LLMs, the payoff average results were normalized between minimum and maximum. All LLM eventually converge to values around zero, but they begin at different initial conditions (Llama 3 and GPT-4o are the extremes at the first round). Claude 3.5 Sonnet converges rapidly to stable payoff values within a few rounds. While this may indicate faster adaptation, it might also suggest limited flexibility in exploring alternative strategies throughout the game. Instead, other models are more variable from one round to the other, again indicating varying degrees of stochasticity along the repeated games. The general downward trend in penalties over rounds for Claude 3.5, Llama 3.1 405B, and Mistral Large indicate progressively increasing mutual cooperation among agents; this is consistent with the strategies traditionally observed in repeated games, where agents reciprocate cooperation to maximize long-term payoffs (Wang et al., 2015). Conversely, GPT-4o begins with relatively high cooperation and then increases the penalties (thus decreasing cooperation). This reflects potential biases toward cooperative behaviors in the case of one-shot Prisoner's Dilemma game (at round one), eventually balanced by context-dependent strategic adaptation. With these results, we thus observe that agents perform behaviors on top of what is purely predicted by the payoff matrix, and that repeated interactions yield different results than the one-shot counterparts.

What is qualitatively described above is quantitatively captured in Figure 5, which summarizes the metrics used to measure, for each LLM, the variability across repeated experiments, inconsistencies across languages, and variability during repeated games (see Section 2.2.4). Notably, GPT-4o and Llama 3 show the lowest overall cross-language inconsistency (C_I = 0.37 and C_I = 0.42 respectively), while Claude 3.5 exhibits the highest C_I (0.79), suggesting a higher sensitivity to prompting language. Moreover, we recognize the higher variability of Claude 3.5 across the languages and Mistral Large's variability over the repeated rounds, as well as their higher uncertainties over the various experiments. As in the zero-sum game, these patterns are stable across the 10 repetitions: the qualitative ordering of the four models in terms of C_I and variability is preserved from run to run, and the credible intervals around the metrics remain narrow compared

with the differences between models. Conversely, GPT-4o and Llama 3 show more consistent results, indicating some stabilizing effect that somehow copes with their stochastic behavior.

4 Discussion

In this work, we examined the strategic behavior of four state-of-the-art LLMs across five languages in two canonical cybersecurity-motivated game-theoretic scenarios, revealing systematic stability differences and notable cross-linguistic effects.

Real-world cyber systems are characterized by high complexity (e.g., partial information or resources, adaptive infiltration schemes, uncertainties) that may divert agents to always perform best-payoff actions. Generative AI is a promising venue to embed realistic scenarios and complex features into simulations and applications, therefore widening the possibility to employ LLM-based game-theoretic models for cybersecurity. However, as LLMs are emerging technologies with unpredictable and often un-interpretable capabilities, it is imperative to systematically assess their capabilities and behaviors. This study provides evidence that LLM agents may behave sub-optimally in key games used for cybersecurity applications, highlighting that the language used for prompting the models, as well as additional traits such as completeness of information or the assigned digital personality of agents may introduce behavioral biases that affect their decision-making during the games. These behavioral differences are consistent with broader observations on contextual and language-dependent biases in LLMs (e.g., Lorè and Heydari, 2024). In cybersecurity settings, this has a concrete operational implication: the strategic responses generated by an LLM-based assistant may shift when deployed in different linguistic environments or with specific features. Organizations and SOC (Security Operations Center) teams should therefore validate LLM-driven decision-support tools in the specific languages and cultural contexts in which they are intended to operate, rather than extrapolating conclusions from English-only evaluations.

Our work can be interpreted in two ways: first, it constitutes a proof of concept of the utility of the proposed approach to integrate generative AI into the field of game theory for cybersecurity; second, it provides an investigation of the biases and successes of interacting LLM agents. Despite being limited to two classes of 2×2 games, based on simplified assumptions that allowed the comparison of outcomes stemming from various bias sources, our study already recognizes several sources of ambiguity in LLM responses, paving the way to future studies focused on specific applications and mitigation of LLM issues. We recognize that, in spite of being canonical, the games considered in this work do not fully capture the complexity and breadth of real-world cybersecurity scenarios. These simplified interactions do not represent, e.g., multi-stage intrusion chains, asymmetric or partial-information settings, stochastic attack surfaces, or networked multi-agent cyber conflicts that characterize operational environments. Additionally, the selection of five languages covers several major linguistic families, but does not exhaust the full spectrum of cultural and linguistic variation. Moreover, the experiments were conducted in simulation without real-world network deployments or adversarial environments, leaving open the question of how these models would perform in operational cybersecurity settings. These relevant questions may constitute basis for future work. Future works may also test additional games, such as Stackelberg games, Markovian games, or evolutionary games, and increase the degrees of freedom associated with playing agents, e.g., by equipping them with complex personalities or different degrees of information, as well as consider multi-agent games on networks.

From the game-theoretical perspective, we have considered games with well-defined equilibrium solutions, namely, minimax strategies in the zero-sum interaction and mutual defection in the repeated Prisoner's Dilemma. Our results show that LLM agents deviate systematically from these equilibria, reflecting their lack of explicit optimisation over payoff functions. Because current LLMs cannot be assumed to follow best-response dynamics, classical convergence guarantees do not apply. Our findings align with prior evidence that current LLMs frequently diverge from game-theoretic equilibria and exhibit context-dependent strategic biases (Lorè and Heydari, 2024; Duan et al., 2024; Fan et al., 2024; Herr et al., 2024). These deviations are also consistent with broader meta-game-theoretic analyses in cybersecurity, such as Yang et al. (2024), which highlight that real-world cyber conflicts often depart from equilibrium predictions due to bounded rationality, asymmetric information and multi-level strategic interactions. By instantiating FAIRGAME within cybersecurity-motivated versions of the zero-sum game and the repeated information-sharing dilemma, we extend these observations to scenarios that more closely reflect operational cyber-defense settings.

In these contexts, sub-optimality may serve as an interpretative asset: LLM agents may be more appropriate for modeling human-like, non-optimal adversarial behavior or for generating exploratory "what-if" simulations, rather than for stand-alone optimisation in high-stakes defense systems. This form of bounded rationality is valuable for modeling human attackers, whose behavior often deviates from perfect rationality. The five-language sample adopted here provides an informative but non-exhaustive view of linguistic sensitivity. Our findings should therefore be interpreted as evidence of cross-language effects rather than as a complete typological analysis; broader multilingual evaluations remain an important direction for future work. Future work may also combine empirical FAIRGAME-style evaluations with analytical tools such as deviation-from-equilibrium measures, stability bounds, or policy-induction analyses to better characterize how LLM-driven strategies relate to normative game-theoretic predictions. These behavioral patterns have direct implications for real cybersecurity operations. For example, an LLM-based assistant that exhibits a strong cooperative bias in a threat-intelligence-sharing game may encourage defenders to share more information, potentially strengthening collective defense, but may also underestimate the risks posed by malicious or opportunistic partners. Conversely, in intrusion-detection scenarios, a model that implicitly favors defection (a pessimistic stance) may overestimate attack likelihood and increase false-positive rates. Understanding the direction and magnitude of these deviations from game-theoretic optima is therefore essential when deciding whether an LLM is best suited for creative scenario exploration, training and education, or as

a decision-support component in high-stakes SOC workflows. Simulating attacker and defender behaviors with AI-driven agents may thus enable better preparation and defense mechanisms, but it also opens the door to malicious uses, such as automated vulnerability discovery or adversarial prompt engineering.

From an operational perspective, the stability profiles identified in this work can be interpreted in terms of concrete SOC workflows. Models exhibiting lower Internal Variability (I_V) and Variability over Rounds (V_R) are better suited for tasks requiring repeatable and dependable recommendations, such as generating incident-response playbooks, suggesting SIEM configuration adjustments, or supporting routine triage. Conversely, models with higher variability may be more suitable for red-teaming and threat-hunting simulations, where diverse trajectories and exploratory what-if scenarios are desirable. Game-theoretic approaches are already used in deployed cybersecurity tools, most notably in resource-allocation and patrol-scheduling systems (Tambe et al., 2012): our results suggest that, before integrating LLM-driven agents into analogous pipelines, practitioners should evaluate candidate models under the relevant game settings and prompting languages and prioritize those demonstrating higher stability and lower cross-language inconsistency before integration in operational SOC pipelines.

Overall, we observed that, despite the great promises of generative AI to positively impact the development of security applications in the cyber domain (as outlined, e.g., by He et al., 2025 when implementing robust mobile networking), LLMs still face notable limitations in handling uncertainty, strategic planning capabilities, and sensitivity to embedded biases. Our methodology and case studies suggest that, before being routinely applied, generative algorithms should be carefully tested by the community in a variety of scenarios and by considering numerous features. Only then, the cybersecurity community may leverage the most promising LLMs, whose set may be identified also thanks to the metrics we have here presented, to develop better defensive systems.

Taken together, these results suggest that current LLM-based agents are not fit-for-purpose as stand-alone optimisers in security-critical systems, particularly where real-time guarantees and strict SLAs are required. Their systematic deviations from equilibrium behavior and their sensitivity to prompt language indicate that rule-based or reinforcement-learning agents remain preferable whenever optimality and predictability are paramount. By contrast, LLMs may be more appropriate as tools for modeling human-like strategic behavior, generating plausible attack or defense narratives, and supporting analysts in exploring what-if scenarios under uncertainty. Our analysis thus complements, rather than replaces, traditional game-theoretic approaches to cybersecurity. Beyond these behavioral observations, several conceptual issues also arise when comparing LLM-driven dynamics to established game-theoretic principles.

Based on these observations, we summarize several practical guidelines for the use of LLM-based agents in cybersecurity workflows. First, for tasks requiring stability and reproducibility, such as generating incident-response playbooks, triage templates, or configuration recommendations, models showing lower I_V and V_R in our experiments (e.g., GPT-4o and Llama-3-405B Instruct) should be preferred. Second, for exploratory or adversarial tasks such as red-teaming and threat-hunting simulations, models exhibiting higher variability (e.g., Claude 3.5 Sonnet or Mistral Large) may be advantageous, provided that outputs remain under human oversight. Third, in multilingual or geographically distributed deployments, organizations should evaluate model behavior in each operational language, in line with prior evidence on cross-linguistic bias and risk assessment in AI systems (Lorè and Heydari, 2024; Gennari et al., 2024). Finally, consistent with meta-game-theoretic perspectives on cyber-defense (Yang and Zhu, 2025), we recommend positioning LLM agents as complementary decision-support or scenario-generation tools, rather than autonomous optimisers in high-stakes defense systems.

As such, we advocate for responsible experimentation frameworks and transparency in reporting LLM-driven cybersecurity simulations. In fact, our case studies point to potential vulnerabilities that need to be carefully considered: if used maliciously, LLMs may behave differently from other traditional algorithms (for instance, by altering cooperative behaviors depending on the language) and bypass solutions tested on more traditional scenarios. This observation thus calls for renewed attention toward these emerging technologies, and suggests the use of coherent testing frameworks, such as FAIRGAME, to systematically test scenarios of increasing complexity. Overall, such tests would enrich our understanding of LLM behaviors in the cyber systems and would help make better predictions and interventions to navigate the newest technologies.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/alessiobuscemi/cybersecurity.

Author contributions

DP: Conceptualization, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. AB: Conceptualization, Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. AD: Methodology, Validation, Writing – original draft, Writing – review & editing. TH: Methodology, Validation, Supervision, Writing – original draft, Writing – review & editing. GC: Supervision, Writing – original draft, Writing – review & editing. PL: Methodology, Validation, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. AB is supported by the Citcom.ai, co-funded by EU/Digital Europe and, in Luxembourg, by the Feder. DP is supported by the European Union through the ERC INSPIRE grant (project number 101076926). TH is supported by EPSRC (grant EP/Y00857X/1).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the European Research Council Executive Agency can be held responsible for them.

References

Acquisti, A., and Grossklags, J. (2005). Privacy and rationality in individual decision making. *IEEE Secur. Priv.* 3, 26–33. doi: 10.1109/MSP.2005.22

Akata, E., Schulz, L., Coda-Forno, J., Oh, S. J., Bethge, M., and Schulz, E. (2025). Playing repeated games with large language models. *Nat. Hum. Behav.* 2025, 1–11. doi: 10.1038/s41562-025-02172-y

Alalawi, Z., Bova, P., Cimpeanu, T., Di Stefano, A., Duong, M. H., Domingos, E. F., et al. (2026). Trust ai regulation? discerning users are vital to build trust and effective ai regulation. *Appl. Math. Comput.* 508:129627. doi: 10.1016/j.amc.2025.129627

Amin, S., and Johansson, K. H. (2019). Preface to the focused issue on dynamic games in cyber security. *Dyn. Games Applic.* 9, 881–883. doi: 10.1007/s13235-019-00335-x

Ara, M., Reboredo, H., Ghanem, S. A., and Rodrigues, M. R. (2012). "A zero-sum power allocation game in the parallel gaussian wiretap channel with an unfriendly jammer," in *2012 IEEE ICCS* (IEEE), 60–64. doi: 10.1109/ICCS.2012.6406109

Avinash, A., and Jain, K. (2025). "Evolving strategies: LLMS as game players," in *2025 4th International Conference on Sentiment Analysis and Deep Learning (ICSADL)* (IEEE), 1009–1014. doi: 10.1109/ICSADL65848.2025.10933026

Balabanova, N., Bashir, A., Bova, P., Buscemi, A., Cimpeanu, T., da Fonseca, H. C., et al. (2025). Media and responsible ai governance: a game-theoretic and LLM analysis. *arXiv:2503.09858*.

Bashir, A., Shamszaman, Z. U., Song, Z., and Han, T. A. (2025). Co-evolutionary dynamics of attack and defence in cybersecurity. *arXiv preprint arXiv:2505.19338*.

Brown, G., Kline, J., Thomas, A., Washburn, A., and Wood, K. (2011). A game-theoretic model for defense of an oceanic bastion against submarines. *Milit. Oper. Res.* 16, 25–40. doi: 10.5711/1082598316425

Buscemi, A., and Proverbio, D. (2024). Large language models' detection of political orientation in newspapers. *arXiv preprint arXiv:2406.00018*.

Buscemi, A., Proverbio, D., Bova, P., Balabanova, N., Bashir, A., Cimpeanu, T., et al. (2025a). Do LLMs trust AI regulation? Emerging behaviour of game-theoretic LLM agents. *arXiv:2504.08640*.

Buscemi, A., Proverbio, D., Di Stefano, A., Han, T. A., Castignani, G., and Lió, P. (2025b). Fairgame: a framework for ai agents bias recognition using game theory. *arXiv preprint arXiv:2504.14325*.

Correia da Fonseca, H., Fernandes, A., Song, Z., Krellner, M., Cimpeanu, T., Balabanova, N., et al. (2025). "Can media act as a soft regulator of safe ai development? a game theoretical analysis," in *ALIFE 2025*.

Do, C. T., Tran, N. H., Hong, C., Kamhoua, C. A., Kwiat, K. A., Blasch, E., et al. (2017). Game theory for cyber security and privacy. *ACM Comput. Surv.* 50, 1–37. doi: 10.1145/3057268

Duan, J., Zhang, R., Diffenderfer, J., Kailkhura, B., Sun, L., Stengel-Eskin, E., et al. (2024). "Gtbench: uncovering the strategic reasoning capabilities of llms via game-theoretic evaluations," in *Advances in Neural Information Processing Systems*, 28219–28253. doi: 10.52202/079017-0885

Etesami, S. R., and Başar, T. (2019). Dynamic games in cyber-physical security: an overview. *Dyn. Games Applic.* 9, 884–913. doi: 10.1007/s13235-018-00291-y

Fan, C., Chen, J., Jin, Y., and He, H. (2024). "Can large language models serve as rational players in game theory? A systematic analysis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 17960–17967. doi: 10.1609/aaai.v38i16.29751

Ferrag, M. A., Alwahedi, F., Battah, A., Cherif, B., Mechri, A., and Tihanyi, N. (2024). Generative ai and large language models for cyber security: All insights you need. *Available at SSRN 4853709*. doi: 10.2139/ssrn.4853709

Fontana, N., Pierri, F., and Aiello, L. M. (2024). Nicer than humans: how do large language models behave in the prisoner's dilemma? *arXiv:2406.13605*.

García, J., and Van Veelen, M. (2018). No strategy can win in the repeated prisoner's dilemma: linking game theory and computer simulations. *Front. Robot. AI* 5:102. doi: 10.3389/frobt.2018.00102

Gennari, J., Lau, S.-,h., Perl, S., Parish, J., and Sastry, G. (2024). "Considerations for evaluating large language models for cybersecurity tasks," in *SEI Insights*, 20.

Hammond, L., Chan, A., Clifton, J., Hoelscher-Obermaier, J., Khan, A., McLean, E., et al. (2025). Multi-agent risks from advanced AI. *arXiv preprint arXiv:2502.14143*.

Han, S., Zhang, Q., Yao, Y., Jin, W., Xu, Z., and He, C. (2024). LLM multi-agent systems: challenges and open problems. *arXiv preprint arXiv:2402.03578*.

Han, T. A., Pereira, L. M., Santos, F. C., Lenaerts, T., et al. (2020). To regulate or not: a social dynamics analysis of an idealised ai race. *J. Artif. Intell. Res.* 69, 881–921. doi: 10.1613/jair.1.12225

Hausken, K. (2002). Probabilistic risk analysis and game theory. *Risk Anal.* 22, 17–27. doi: 10.1111/0272-4332.t01-1-00002

Hausken, K. (2020). Cyber resilience in firms, organizations and societies. *Internet Things* 11:100204. doi: 10.1016/j.iot.2020.100204

Hausken, K. (2024). Fifty years of operations research in defense. *Eur. J. Oper. Res.* 318, 355–368. doi: 10.1016/j.ejor.2023.12.023

Hausken, K., Welburn, J. W., and Zhuang, J. (2024). A review of attacker-defender games and cyber security. *Games* 15:28. doi: 10.3390/g15040028

He, L., Sun, G., Niyato, D., Du, H., Mei, F., Kang, J., et al. (2025). Generative ai for game theory-based mobile networking. *IEEE Wirel. Commun.* 32, 122–130. doi: 10.1109/MWC.007.2400133

Herr, N., Acero, F., Raileanu, R., Perez-Ortiz, M., and Li, Z. (2024). "Large language models are bad game theoretic reasoners: Evaluating performance and bias in two-player non-zero-sum games," in *ICML 2024 Workshop on LLMs and Cognition*.

Huang, J.-,t., Li, E. J., Lam, M. H., Liang, T., Wang, W., Yuan, Y., et al. (2025). "Competing large language models in multi-agent gaming environments," in *13th International Conference on Learning Representations*.

Ji, Z., Yu, W., and Liu, K. R. (2010). A belief evaluation framework in autonomous manets under noisy and imperfect observation: Vulnerability analysis and cooperation enforcement. *IEEE Trans. Mobile Comput.* 9, 1242–1254. doi: 10.1109/TMC.2010.87

Jia, J., Yuan, Z., Pan, J., McNamara, P. E., and Chen, D. (2025). Large language model strategic reasoning evaluation through behavioral game theory. *arXiv preprint arXiv:2502.20432*.

Kamhoua, C. A., Pissinou, N., and Makki, S. K. (2010). "Game theoretic analysis of cooperation in autonomous multi hop networks: The consequences of unequal traffic load," in *2010 IEEE Globecom Workshops* (IEEE), 1973–1978. doi: 10.1109/GLOCOMW.2010.5700289

Kamhoua, C. A., Rodriguez, M., and Kwiat, K. A. (2014). "Testing for hardware trojans: a game-theoretic approach," in *International Conference on Decision and Game Theory for Security* (Springer), 360–369. doi: 10.1007/978-3-319-12601-2_22

Kasri, W., Himeur, Y., Alkhazaleh, H. A., Tarapiah, S., Atalla, S., Mansoor, W., et al. (2025). From vulnerability to defense: the role of large language models in enhancing cybersecurity. *Computation* 13:30. doi: 10.3390/computation13020030

Kostyuk, N. (2013). "The digital prisoner's dilemma: challenges and opportunities for cooperation," in *2013 World Cyberspace Cooperation Summit IV (WCC4)*, 1–6. doi: 10.1109/WCS.2013.7050508

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., and Zou, J. (2023). GPT detectors are biased against non-native english writers. *Patterns* 4:100779. doi: 10.1016/j.patter.2023.100779

Liu, Y., Feng, D., Lian, Y., Chen, K., and Zhang, Y. (2013). "Optimal defense strategies for ddos defender using bayesian game model," in *9th Conference on Information Security Practice and Experience (ISPEC)* (Springer), 44–59. doi: 10.1007/978-3-642-38033-4_4

Loré, N., and Heydari, B. (2024). Strategic behavior of large language models and the role of game structure vs. contextual framing. *Sci. Rep*. 14:18490. doi: 10.1038/s41598-024-69032-z

Lu, Y., Aleta, A., Du, C., Shi, L., and Moreno, Y. (2024). LLMS and generative agent-based models for complex systems research. *Phys. Life Rev*. 51, 283–293. doi: 10.1016/j.plrev.2024.10.013

McKelvey, R. D., and Palfrey, T. R. (1998). Quantal response equilibria for extensive form games. *Exper. Econ*. 1, 9–41. doi: 10.1023/A:1009905800005

Morgan, S. (2020). Cybercrime to cost the world $10.5 trillion annually by 2025. Cybercrime Magazine.

Nguyen, A. T., Anand, S. C., and Teixeira, A. M. (2022). "A zero-sum game framework for optimal sensor placement in uncertain networked control systems under cyber-attacks," in *2022 IEEE 61st Conference on Decision and Control (CDC)* (IEEE), 6126–6133. doi: 10.1109/CDC51059.2022.9992468

Owen, G. (2013). *Game Theory*. Leeds: Emerald Group Publishing.

Pala, A., and Zhuang, J. (2019). Information sharing in cybersecurity: a review. *Dec. Anal*. 16, 172–196. doi: 10.1287/deca.2018.0387

Patel, N., and Trivedi, S. (2020). Leveraging predictive modeling, machine learning personalization, nlp customer support, and ai chatbots to increase customer loyalty. *Empir. Quests Manage. Essenc*. 3, 1–24.

Petrosyan, A. (2024). *Estimated cost of cybercrime worldwide 2018–2029*. Available online at: https://www.statista.com/forecasts/1280009/cost-cybercrime-worldwide (Accessed June 10, 2025).

Schoenherr, J. R., and Thomson, R. (2020). "Beyond the prisoner's dilemma: the social dilemmas of cybersecurity," in *2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)* (IEEE), 1–7. doi: 10.1109/CyberSA49311.2020.9139644

Shiva, S., Roy, S., and Dasgupta, D. (2010). "Game theory for cyber security," in *Proceedings of the Sixth Annual Workshop on Cyber Security and Information Intelligence Research*, 1–4. doi: 10.1145/1852666.1852704

Shukla, P., An, L., Chakrabortty, A., and Duel-Hallen, A. (2022). A robust stackelberg game for cyber-security investment in networked control systems. *IEEE Trans. Control Syst. Technol*. 31, 856–871. doi: 10.1109/TCST.2022.3207671

Spyridopoulos, T., Karanikas, G., Tryfonas, T., and Oikonomou, G. (2013). A game theoretic defence framework against dos/ddos cyber attacks. *Comput. Secur*. 38, 39–50. doi: 10.1016/j.cose.2013.03.014

Srinivasan, V., Nuggehalli, P., Chiasserini, C.-F., and Rao, R. R. (2003). "Cooperation in wireless ad hoc networks," in *IEEE INFOCOM 2003* (IEEE), 808–817. doi: 10.1109/INFCOM.2003.1208918

Sun, H., Wu, Y., Cheng, Y., and Chu, X. (2025). Game theory meets large language models: a systematic survey. *arXiv preprint arXiv:2502.09053*.

Swoopes, C., Holloway, T., and Glassman, E. L. (2025). The impact of revealing large language model stochasticity on trust, reliability, and anthropomorphization. *arXiv preprint arXiv:2503.16114*.

Tambe, M., Jain, M., Pita, J. A., and Jiang, A. X. (2012). "Game theory for security: key algorithmic principles, deployed systems, lessons learned," in *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)* (IEEE), 1822–1829. doi: 10.1109/Allerton.2012.6483443

Tessler, M. H., Bakker, M. A., Jarrett, D., Sheahan, H., Chadwick, M. J., et al. (2024). AI can help humans find common ground in democratic deliberation. *Science* 386:eadq2852. doi: 10.1126/science.adq2852

Von Neumann, J., and Morgenstern, O. (2007). "Theory of games and economic behavior: 60th anniversary commemorative edition," in *Theory of Games and Economic Behavior*. Princeton: Princeton University Press. doi: 10.1515/9781400829460

Wang, Y., Wang, Y., Liu, J., Huang, Z., and Xie, P. (2016). "A survey of game theoretic methods for cyber security," in *2016 IEEE 1st International Conferences on Data Science in Cyberspace (DSC)* (IEEE), 631–636. doi: 10.1109/DSC.2016.90

Wang, Z., Kokubo, S., Jusup, M., and Tanimoto, J. (2015). Universal scaling for the dilemma strength in evolutionary games. *Phys. Life Rev*. 14, 1–30. doi: 10.1016/j.plrev.2015.04.033

Wang, Z., Song, R., Shen, C., Yin, S., Song, Z., Battu, B., et al. (2024). Large language models overcome the machine penalty when acting fairly but not when acting selfishly or altruistically. *arXiv:2410.03724*.

Wellman, M. P., Tuyls, K., and Greenwald, A. (2025). Empirical game theoretic analysis: a survey. *J. Artif. Intell. Res*. 82, 1017–1076. doi: 10.1613/jair.1.16146

Xiao, Y., Shi, G., and Zhang, P. (2025). Towards agentic ai networking in 6G: a generative foundation model-as-agent approach. *arXiv preprint arXiv:2503.15764*.

Yamin, M. M., Hashmi, E., Ullah, M., and Katt, B. (2024). Applications of LLMS for generating cyber security exercise scenarios. *IEEE Access* 12, 143806–143822. doi: 10.1109/ACCESS.2024.3468914

Yang, Y., Du, H., Sun, G., Xiong, Z., Niyato, D., and Han, Z. (2024). Exploring equilibrium strategies in network games with generative AI. *IEEE Netw*. 39, 191–200. doi: 10.1109/MNET.2024.3521887

Yang, Y.-T., and Zhu, Q. (2025). Toward a multi-echelon cyber warfare theory: a meta-game-theoretic paradigm for defense and dominance. *arXiv preprint arXiv:2509.08976*.

Zhang, J., Bu, H., Wen, H., Liu, Y., Fei, H., Xi, R., et al. (2025). When llms meet cybersecurity: a systematic literature review. *Cybersecurity* 8, 1–41. doi: 10.1186/s42400-025-00361-w

Zhu, Q., and Başar, T. (2011). "Robust and resilient control design for cyber-physical systems with an application to power systems," in *2011 50th IEEE Conference on Decision and Control and European Control Conference* (IEEE), 4066–4071. doi: 10.1109/CDC.2011.6161031