Riddle Me This: Evaluating Large Language Models in Solving Word-Based Games

Raffaele Manna, Maria Pia di Buono, Johanna Monti
UniOr NLP Research Group
University of Naples "L'Orientale", Italy
{raffaele.manna, mpdibuono, jmonti}@unior.it

Abstract

In this contribution, we examine the proficiency of Large Language Models (LLMs) in solving the linguistic game "La Ghigliottina," the final game of the popular Italian TV quiz show "L'Eredità". This game is particularly challenging as it requires LLMs to engage in semantic inference reasoning for identifying the solutions of the game. Our experiment draws inspiration from Ghigliottin-AI, a task of EVALITA 2020, an evaluation campaign focusing on Natural Language Processing (NLP) and speech tools designed for the Italian language. To benchmark our experiment, we use the results of the most successful artificial player in this task, namely Il Mago della Ghigliottina. The paper describes the experimental setting and the results which show that LLMs perform poorly.

Keywords: Large language model, Ghigliottin-AI, Word-based Games

1. Introduction

Researchers in Artificial Intelligence (AI) and Natural Language Processing (NLP) have shown interest in Language games, which derive their challenge and excitement from the complexity and ambiguity of natural language. A particular challenging language game is "La Ghigliottina", the final game of the popular Italian TV quiz show "L'Eredità". The game involves a single player, who is given a set of five words (clues), unrelated one to each other, but related with a sixth word that represents the solution to the game. In 2020 EVALITA, a recurring evaluation campaign focusing on NLP and speech tools designed for the Italian language, proposed the Ghigliottin-AI task (Basile et al., 2020) to assess artificial agents in the solution of "La Ghigliottina". Participants in Ghigliottin-AI are asked with developing an artificial player capable of solving the linguistic challenges presented in the game "La Ghigliottina". In the aftermath of the Ghigliottin-AI task, this contribution aims to examine the ability of cutting-edge Large Language Models in solving the Ghigliottina game, which involves inferring the solution through identifying the hidden semantic connections with the provided clues. This paper is organized as follows: in Section 2 we briefly present the use of games in testing the reasoning and inference abilities of NLP and AI systems. In Section 3 we present the Ghigliottin-AI task and the results obtained by the artificial players that took part in the task. In Section 4 we provide all the information (data, LLM models and prompts) concerning our experimental settings to evaluate the abilities of different LLMs in solving the GhigliottinAI language game. Discussion of results is presented in Section 5. Conclusions are in Section 6.

2. Related Work

In this section, we briefly survey the use of games as a means to assess the efficacy of NLP tools in problem-solving tasks. Some achievements in artificial intelligence are linked to games such as for instance *Jeopardy*, where contestants respond to clues in the form of answers by phrasing their replies as questions. In 2011, IBM's Watson DeepQA computer defeated the show's two foremost all-time champions of this game (Ferrucci et al., 2013). In particular, language games, such as the *Wheel of Fortune* or *Who Wants to be a Millionaire?* (Lam et al., 2012) (Molino et al., 2015), have been used as means to assess the capabilities of NLP and AI systems, as they provide an interesting and challenging playground to evaluate their reasoning and inference capabilities (Yannakakis and Togelius, 2018). Another particularly appealing game is solving crossword puzzles. A first attempt is *Proverb* (Littman et al., 2002), which leverages extensive repositories containing clues and solutions to past crossword puzzles. *WebCrow* (Ernandes et al., 2008), the first solver for Italian crosswords, instead, relies mainly on information sourced from the Web, and a set of previously solved games.

As mentioned in the Introduction the *Ghigliottina* game is particularly challenging and has inspired various scholars in solving it. In (Semeraro et al., 2012) and (Basile et al., 2014), the authors present OTTHO (On the Tip of my THOught), an artificial player for the Guillotine game. OTTHO is based on a knowledge infusion procedure that uses NLP techniques to analyze unstructured data from open web sources like Wikipedia, creating a repository of linguistic competencies and factual knowledge. In 2018 the *Mago della Ghigliottina* (Sangati et al., 2018) participated as UNIOR4NLP for the first time in the shared task NLP4FUN (Basile et al., 2018), which was part of the EVALITA 2018, a periodic evaluation campaign of NLP and speech tools for the Italian language. The system, available also as a Telegram bot,[1] relies on linguistic resources and artificial intelligence and achieves better results than human players. In addition to solving a game, *Mago della Ghigliottina* can also generate new game instances and challenge the users to match the solution. The *Mago della Ghigliottina* took part in the new edition of the NLP4FUN task, titled Ghigliottin-AI, resulting again as the best artificial player, outperforming human players and competitor artificial players (see Section 3). Recently LLMs were tested in solving *Wordle*,[2] a game owned by the New York Times, where players have six attempts to guess a five-letter word. The experiment showed that LLMs lack the inference skills needed to solve the game.

3. GhigliottinAI

As part of EVALITA 2020, the Ghigliottin-AI[3] task was organised, a new edition of the NLP4FUN task proposed in EVALITA 2018 (Basile et al., 2018), aimed at the realisation of an open competition between Artificial Intelligence (AI) systems to solve the game “La Ghigliottina". The Ghigliottin-AI task is inspired by the final game of the Italian TV show “L'Eredità". This game was chosen because it represents a very interesting test bed for AI systems focused on semantic aspects of natural language: the solution of the language game is based on the semantic relationships existing between each of the five proposed clues and the solution word. For example, given the set of Italian clues *conoscere* (to know), *grado* (degree), *modello* (model), *ideale* (ideal) and *divina* (divine) the solution is *perfezione* (perfection) because this word relates to the clues in the following way: *conoscere alla perfezione* (to perfectly know), *grado di perfezione* (degree of perfection), *modello di perfezione* (model of perfection), *ideale di perfezione* (ideal of perfection) and *perfezione divina* (divine perfection).

The underlying idea of the Ghigliottin-AI task was that artificial players for that game could take advantage from the availability of open repositories on the Web, such as Wikipedia, that provide the system with the cultural and linguistic background needed to understand clues (Basile et al., 2014; Semeraro et al., 2009, 2012). Before the competition, a set of 300 instances of the game together with their solution taken from the last editions of the TV game were provided to developers in a JSON format as training data for their players. The evaluation was carried out using a Remote Evaluation Server (RES) named Ghigliottiniamo[4], which facilitated real-time submission of solutions by both human participants and artificial systems (bots) to the TV game. Ghigliottiniamo randomly provided the test set at intervals, presenting a single game challenge to registered systems. The RES imposed a time constraint, similar to the original TV game, allowing systems to submit a single solution within 60 seconds from the challenge. Solutions received after this time frame were discarded, mirroring the time-sensitive nature of the original game. This protocol was applied consistently in evaluating systems participating in Ghigliottin-AI. Two teams participated to the competition: *Mago della Ghigliottina* (Sangati et al., 2020) and GUL.LE.VER (De Francesco, 2020).

Mago della Ghigliottina is based on the analysis of real game instances. As highlighted by the authors (Sangati et al., 2020), game instances indicate that connections between clues and solution pertain to a specific linguistic phenomenon, namely Multiword Expression (MWE)(Sag et al., 2002; Constant et al., 2017). A MWE is a sequence of words that presents some characteristic behaviour (at the lexical, syntactic, semantic, pragmatic or statistical level) and whose interpretation crosses the boundaries between words. During the analysis six patterns that identify MWEs connecting clue/solution pairs were identified:

- **A-B (Noun-Adjective, Adjective-Noun, Verb-Noun, Noun-Noun)**: *permesso premio* (‘permit price’ → good behaviour license)
- **A-determiner-B**:*dare il permesso* (‘give the permit’ → authorize)];
- **A-conjunction-B**: *stima e affetto* (esteem and affection);
- **A-preposition-B**: *colpo di coda* (‘flick of tail’ → last ditch effort);
- **A-articulated preposition-B** : *virtù dei forti* , part of the famous Italian proverb La calma è la virtù dei forti (patience is the virtue of the strong);
- **A+B**: compounds such as radio + attivita = *radioattivita‘* (radio + activity = radioactivity).

[1] https://t.me/Unior4NLPbot

[2] https://github.com/manavgup/wordle-llm-solver

[3] https://ghigliottin-ai.github.io/

[4] https://quiztime.net/web/#!/leaderboard1

Therefore, *Mago della Ghigliottina* explores word co-occurrencence in frequent collocations or idioms, word similarity or word relatedness as a basis of the semantic relationship of clues and solutions in a number of freely large available corpora, such as Paisà[5], itWaC[6], Wikiquote[7] and other linguistic resources. *Mago della Ghigliottina* proved to be the best performing artificial player with an accuracy score of .68.

GUL.LE.VER positioned #2 in the competition, with an accuracy score of .26 and .46 R@10, achieving results comparable to human players of the TV game. This player is based on the Glove vector representation of the words (Pennington et al., 2014) on the basis of a large collected dataset, containing the Italian Wiktionary, Wikiquote, Wikipedia (only titles), the Italian Collocations Dictionary and other resources scraped on the web containing Italian multiword expressions, proverbs and songs titles.

4. Experiment

This section presents the experimental settings to evaluate the abilities of different LLMs in solving GhigliottinAI language game. Section 4.1 presents the data on which LLMs were tested at the GhigliottinAI game, while Section 4.2 describes the LLMs and the parameters used to generate their outputs for each game instance. The outputs of the LLMs at GhigliottinAI were elicited using different prompting techniques. In Section 4.3, the different prompting techniques used are listed and examples of prompts provided to LLMs are shown. Finally, in Section 4.4, the performances obtained by the LLMs in solving GhigliottinAI using the different prompting techniques are shown.

The game instances were solved between mid-December and mid-February. During this time, the two leading AI firms, Google[8] and OpenAI[9], remained active in releasing updated versions of their respective LLMs. As in 4.4 and specifically in section 5, the updates had a notable impact on the performance of the LLMs in solving the GhigliottinAI game instances.

The game instances used to test the LLMs, as well as the solutions generated for each prompting technique, are available in this repository[10].

4.1. Data

We used data from a shared task organized as part of the Italian NLP tools evaluation campaign: Evalita [11]. Following up on Section 3, the 2020 edition of Evalita introduced a shared task named "Solving the Ghigliottina with AI," along with the release of training game instances. The test set consists of 350 game instances [12], released in an excel sheet. The excel sheet is therefore composed of 350 rows representing the games instances and 8 columns. The first column contains the game ID, columns from 2 to 6 represent the clue words, and the last column contains the solution words for each instance. In Figure 1, we provide an example of the excel file with game instances.

We used the game instances contained in the test set to evaluate the performance of the LLMs. This approach allows us to compare the performance of the LLMs to the performance of the automatic solvers presented in the shared tasks discussed in section 3. Also some game instances from the training set of GhigliottinAI were used to provide a game demonstration in some prompts to enable in-context learning (Brown et al., 2020; Min et al., 2022).

4.2. Large Language Models

In an effort to evaluate their aptitude at the GhigliottinAI game, four LLMs, including ChatGPT-3.5, ChatGPT-4, Bard and Gemini-Pro, were systematically exposed to the game. To conduct the experiments, we used Chatbot Arena[13], a benchmark platform that offers access to several LLMs via a web graphical user interface (Zheng et al., 2024).

Despite the prompting technique chosen, we crafted a block of prompts containing a number of game instances considering the maximum sample length for each LLMs. We tested the LLMs on the benchmark platform[14] using configurable parameters like Temperature (set to 0.7), Top P (set to 1), and Max Output Tokens (set to 1024). The aforementioned parameters were configured separately for each LLM.

We define a set of prompts while considering the maximum token length that can be processed by

[5] https://www.corpusitaliano.it/

[6] https://wacky.sslmit.unibo.it/doku.php?id=corpora

[7] https://it.wikiquote.org/wiki/Pagina_principale

[8] https://blog.google/products/gemini/bard-gemini-advanced-app/

[9] https://platform.openai.com/docs/models/continuous-model-upgrades

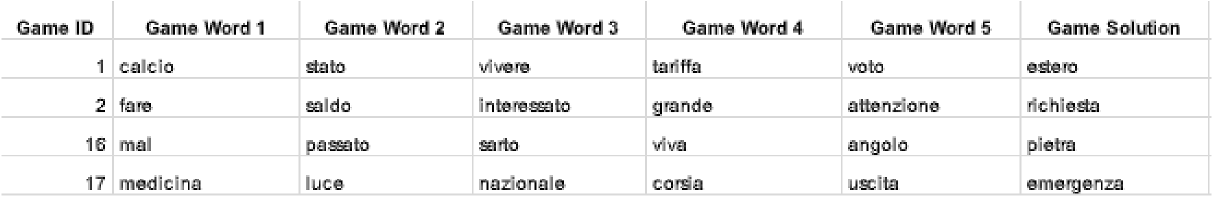
[10] https://github.com/RaffaeleMann/eval_LLMs_ghigliottina

[11] https://ghigliottin-ai.github.io/

[12] The GhigliottinAI test data are not freely accessible and are available only on request.

[13] https://chat.lmsys.org/

[14] More specifically, we used the following LLMs from the platform: GPT-4 (GPT-4-1106-preview), Bard (bard-jan-24-gemini-pro), Gemini-Pro (gemini-pro-dev-api), and GPT-3.5 (GPT-3.5-Turbo-0613).



Game ID	Game Word 1	Game Word 2	Game Word 3	Game Word 4	Game Word 5	Game Solution
1	calcio	stato	vivere	tariffa	voto	estero
2	fare	saldo	interessato	grande	attenzione	richiesta
16	mal	passato	sarto	viva	angolo	pietra
17	medicina	luce	nazionale	corsia	uscita	emergenza

Figure 1: A screenshot of the Excel file containing GhigliottinAI game instances

the LLMs. For instance, given a context size of 8,000 tokens for GPT-4, the block containing our prompts levels out at approximately 20 game instances included.

4.3. Prompts

As far as prompts are concerned, following Wang et al. (2023), we define In-Context Learning (ICL) settings to evaluate LLMs, which include zero- and few-shot approaches.

Zero-Shot Prompting This approach aims to explore how the LLMs handle the task with no prior examples or training, relying solely on their preexisting knowledge and the inherent ability to understand and generate language.

For the zero-shot prompting technique (ZSP), we define two distinct prompts (i.e., ZSP1 and ZSP2), each designed to elicit a different focus from the selected LLMs on the connection established between the clues and the solution. In particular, by implementing these distinct prompts, we intend to assess the versatility of the LLMs in deducing the correct word associations under the constraints of zero-shot learning conditions.

To avoid the presence of extrinsic hallucinations in the results, namely the presence of additional text besides the desired output, we constraint the prompts, phrased in Italian, to force models to provide just the solution for each game. For this reason, we specify that (i) the games provided in the list are independent of each other, (ii) the solution must differ from the words already included in a game, (iii) the answer should not include any additional text but just the solution to each game.

Starting from the list of games, each one composed by a list of clues ([CLUES]), ZSP1 and ZSP2 differ in that the former asks for a 'related word', namely the solution ([SOLUTION]), without specifying the type of existing relationship, the latter specifies that the [SOLUTION] should be 'semantically related' to the [CLUES], as shown below.

- **ZSP-1** *Per ciascun gioco* [CLUES] *in questa lista, scrivi una sola altra parola connessa a ciascuna delle cinque parole incluse in ciascun gioco* (For each game [CLUES] in this list, write only one other word that is related to each of the five words included in each game).

- **ZSP-2** *Per ciascun gioco* [CLUES] *in questa lista, scrivi una sola altra parola semanticamente connessa a ciascuna delle cinque parole incluse in ciascun gioco* (For each game [CLUES] in this list, write only one other word that is semantically related to each of the five words included in each game).

Few-Shot Prompting In the context of Few-Shot Prompting (FSP), two different prompts have been defined. The first one (FSP1) includes one example ([GAME SOLVED]), namely a list of [CLUES] along with the solution , while the second prompt (FSP2) presents three [GAME SOLVED], as it follows:

- **FSP1** *Dato il seguente esempio* [GAME SOLVED], *per ciascun gioco* [CLUES] *in questa lista, scrivi una sola altra parola connessa a ciascuna delle cinque parole incluse in ciascun gioco* (Given the following example [GAME SOLVED], for each game [CLUES] in this list, write only one other word that is related to each of the five words included in each game)

- **FSP2** *Dati i seguenti esempi* [GAME SOLVED], *per ciascun gioco* [CLUES] *in questa lista, scrivi una sola altra parola connessa a ciascuna delle cinque parole incluse in ciascun gioco* (Given the following example [GAME SOLVED], for each game [CLUES] in this list, write only one other word that is related to each of the five words included in each game)

We force the models to return just the [SOLUTION], specifying the same constraints used for ZSP.

Examples With reference to the provided [GAME SOLVED], considering the MWE patterns connecting clue and solution pairs, we manually select from the training set examples which are representative of specific phenomena. For the [GAME SOLVED] provided in FSP1, we choose an example which includes A-B, A-preposition-B and A-articulated preposition-B MWEs, as it follows:

- **Example 1**
  [CLUES]: Nicola, *Roma* (Rome), *farina* (flour),

pranzo (lunch), *poltrona* (armchair)
[SOLUTION]: *sacco* (sack).

Each clue is related to the [SOLUTION] according to the following MWE patterns:

- **A-B pattern**: Nicola Sacco[15]
- **A-preposition-B pattern**: *sacco di farina* (flour bag), *sacco di Roma* (sack of Rome[16]), *poltrona a sacco* (bean bag chair)
- **A-articulated preposition-B pattern**: *pranzo al sacco* (packed lunch)

To run the FSP2 which presents three examples, we add two [GAME SOLVED] whose clue/solution pairs are related by other patterns, as shown below.

- **Example 2**
  [CLUES]: *bello* (nice), *inter*, *vino* (wine), *indosso* (wear), *fronte* (forehaed/front)
  [SOLUTION]: *porto* (port/freight)
- **Example 3**
  [CLUES]: spedito, gigante, uomo, carica, vita
  [SOLUTION]: passo

Specifically, Example 2 shows the following phenomena:

- **A-B pattern**: *Portobello* (the name of an Italian tv show but also a place), *interporto* (freight village)
- **Semantic relations**: hypernymy (*porto* (Port) is a type of wine), synonimy (*porto* and *indosso* may refer to the same meaning to wear)
- **A-articulated preposition-B pattern**: *Fronte del porto*[17] (On the Waterfront), a 1954 movie.

In the third example all the clue/solution pairs are related as they occur together as part of an idiom, namely *a passo spedito* (at a fast pace), *fare passi da gigante* (make great strides), *a passo d'uomo* (at a walking pace), *a passo di carica* (at a charge pace), *passare a miglior vita* (to pass away).

4.4. Results

In this section, we present the results obtained from the four LLMs on the 350 game instances included in the GhigliottinAI test set. We calculate the accuracy as the ratio between solved games on the total games. In Table 1, we show the number of correct solutions together with the accuracy rate for each of the models in both ZSP and FSP settings. The accuracy score is the evaluation metric adopted by Basile et al. (2020) in the original shared task. The results show that the four LLMs performed poorly on both the ZSP and FSP settings.

GPT-4 and Gemini-Pro perform the best in the FSP2 setting when shown three examples of [GAME SOLVED]. Both models achieved an accuracy of .022, which was an improvement over the other settings. In particular, GPT-4 and Gemini-Pro doubled the accuracy scored in both ZSP1 and ZSP2.

Bard also showed efficient in-context learning when given game instances in FSP1 and FSP2. In FSP1, Bard was the best LLM at solving game instances, with an accuracy of 0.14. In FSP2, Bard accuracy was .02, which was slightly worst than the accuracy scored by Gemini-Pro and GPT-4 (i.e., .022).

GPT-3.5, on the other hand, did not seem to benefit from in-context learning. In both FSP1 and FSP2, GPT-3.5 had the lowest accuracy (i.e., .005), proving any improvement in comparison with the results from the ZSP settings. In fact, in the case of GPT-3.5, the accuracy achieved in ZSP1 turns out to be the best performance by this LLM (i.e., .008).

To further evaluate the performance of the LLMs, we also show the number of solutions that they share in each setting. This gives us an idea of how often the LLMs agree on the solution to a game instance. Tables 2 and 3 show the number of (whether correct or not) solutions shared between each pair of LLMs, respectively in ZSP and FSP settings. The highest number of shared solutions for each pair is highlighted in bold. The highest number of shared solutions for each pair is highlighted in bold.

For instance, the GPT family of LLMs from OpenAI share the most solutions for ZSP1 (Table 2), while Google LLMs share the most solutions for both ZSP2 and FSP2 (Table 2 and 3).

For ZSP settings, there is an exception to this trend. Indeed, Gemini-Pro and GPT-3.5, which are from different families, share the most solutions in FSP1 (Table 3).

In this context, one possible explanation for the shared solutions is that the LLMs were trained on similar data sets. This is supported by the fact that the LLMs performed similarly in the ZSP setting, where they were not given any examples of game instances. Another possible explanation is that the LLMs are all using similar in-context learning techniques. This is supported by the fact that the LLMs all improved their performance in the FSP2 setting, where they were given a few and sufficient examples of game instances.

[15] https://en.wikipedia.org/wiki/Sacco_and_Vanzetti

[16] https://en.wikipedia.org/wiki/Sack_of_Rome_(1527)

[17] https://it.wikipedia.org/wiki/Fronte_del_porto

LLM	**ZSP1**		**ZSP2**		**FSP1**		**FSP2**	
	Correct	Acc.	Correct	Acc.	Correct	Acc.	Correct	Acc.
GPT-3.5	3	.008	0	0	2	.005	2	.005
GPT-4	**4**	**.011**	**4**	**.011**	2	.005	**8**	**.022**
Bard	1	.002	2	.005	**5**	.014	7	.02
Gemini-Pro	2	.005	**4**	**.011**	3	.008	**8**	**.022**

Table 1: Number of correct answers and accuracy score for ZSP and FSP

ZSP1	GPT-3.5	GPT-4	Bard	Gemini-Pro	**ZSP2**	GPT-3.5	GPT-4	Bard	Gemini-Pro
GPT-3.5	350	**23**	4	3	GPT-3.5	350	3	7	8
GPT-4	**23**	350	1	2	GPT-4	3	350	12	17
Bard	4	1	350	13	Bard	7	12	350	**25**
Gemini-Pro	3	2	13	350	Gemini-Pro	8	17	**25**	350

Table 2: Shared solutions for ZSP1 and ZSP2

FSP1	GPT-3.5	GPT-4	Bard	Gemini-Pro	**FSP2**	GPT-3.5	GPT-4	Bard	Gemini-Pro
GPT-3.5	350	3	6	**37**	GPT-3.5	350	10	7	19
GPT-4	3	350	17	9	GPT-4	10	350	29	17
Bard	6	17	350	12	Bard	7	29	350	**31**
Gemini-Pro	**37**	9	12	350	Gemini-Pro	19	17	**31**	350

Table 3: Shared solutions for FSP1 and FSP2

5. Discussion

In this section, we present an in-depth result analysis to provide some insights of the semantic inference capabilities of LLMs.
As far as the results are concerned, we notice the presence of shared characteristics among these, in that we can identify different types of incorrect answers:

- **Complete clue overlapping** In some cases the proposed [SOLUTION] overlaps with a word in the [CLUES]. For instance, in ZSF1 Bard presents a high number of overlapping solutions, as in ID 2 when the model answers *saldo* (discount) that is also one of the [CLUES].

- **Partial clue overlapping** These results refer to solutions which are derived from one of the [CLUES], e.g. a noun from a verb, as in ZSP1 for ID 266, when GPT-3.5 provides the [SOLUTION] *conteggio* (count) and the first clue is *contare* (to count).

- **Semantic relatedness** These answers usually are generated leveraging the taxonomic relations of one of the [CLUES], so that they are semantically related to one of the [CLUES] and/or to the [SOLUTION]. For instance, in the game ID 167, the solution proposed by GPT-3.5 is *sentimento* (feeling), as one of the [CLUES] is *amore* (love), and the correct [SOLUTION] is *odio* (hate).

- **Clue synonymy** In some cases, the models propose a synonym of one of the [CLUES]. For instance, in ZSP1 to ID 204 (further discussed later), GPT-3.5 answers *guardia* (watchman), a synonym of *custode*, which is presented in the [CLUES] for that game.
Similarly, in ID 169, GPT-3.5 presents the [SOLUTION] *abitazione* (home), as one of the [CLUES] is the synonym *casa*, while the correct answer is *strada* (road).

- **Clue interference** For some of the games, there is a clue interference that is probably related to the fact that the association between one of the [CLUES] and the possible answer is stronger than others. For instance, in the game ID 69 to the ZSP2 prompt, all the models answer *deserto* (desert), as one of the clues is Sahara.

Considering all the proposed games, the highest agreement among models, on a correct [SOLUTION], that means three models out of four guess the [SOLUTION], happens only in two cases. In ZSP1 setting, this is the game ID 349, shown below.

- **ID 349**
[CLUES]: *coperto* (covered), *compagnia* (company), *auto* (carro), *agente* (agent), *vita* (life)
[SOLUTION]: *assicurazione* (insurance).

Each clue is related to the solution according to (i) the A-preposition-B pattern, i.e., *coperto da assicurazione* (covered by insurance), *compagnia di assicurazione* (insurance company), *agente di assicurazioni*[18] (insurance agent); (ii) the A-B pattern, i.e., *assicurazione auto* (car insurance); (iii) the A-articulated preposition-B pattern, i.e., *assicurazione sulla vita* (life insurance).

GPT-4 does not solve the aforementioned game, as the proposed [SOLUTION] is *musica* (music). This could be the results of an interference from two of the clues, namely *compagnia* (company) and *agente* (manager), which occur in MWEs as *compagnia musicale* (music company) and *agente musicale* (music manager) respectively. The other case of highest agreement happens on the game ID 153 when we provide three examples in FSP2.

- **ID 153**
  [CLUES]: *lavare* (to wash), *nuovo* (new), *espressione* (look), *maschera* (mask), *pallido* (pallid)
  [SOLUTION]: *viso* (face)

In such case, Gemini-Pro disagrees on the answer and provide the [SOLUTION] *sapone* (soap), due to the presence of the verb *lavare* (to wash) as first clue which presumably causes an interference on the provided solution.

In only one case we have the full agreement that is when all the models propose the same answer. This is the case of the game ID 69 in the ZSP2 setting, when the models agree on the incorrect [SOLUTION], *deserto* (desert), due to an interference from the clue Sahara.

To further evaluate the results, we propose a comparative analysis for each of the models.

GPT-3.5 In ZSP1, we notice that GPT-3.5 identifies the solution in the game below.

- **ID 41**
  [CLUES]: *nazionale* (national), *muscolo* (muscle), *lavoro* (job), *proposta* (proposal), *firmare* (to sign)
  [SOLUTION]: *contratto* (contract)

Due to the presence of (i) A-B pattern: *contratto nazionale* (national contract), *muscolo contratto* (contracted muscle); (ii) A-preposition-B pattern: *contratto di lavoro* (employment contract), *proposta di contratto* (contract proposal); (iii) idiom: *firmare un contratto* (sign a contract).

The additional specification about the semantic relatedness in ZSP2 worsens the results, as GPT-3.5 fails all the games, including the game ID 41. Indeed, the proposed [SOLUTION] to the ZSP2 for this game is *fede* (faith). Our hypothesis is that this result is affected by the first clue *nazionale*, as there exist some books whose title contains both the words and also some conservative political parties refer to *fede* and *nazione* to support their ideologies.

Similarly, in the FSP1 setting, the answer to ID 41 is *strada* (road), as in *strada nazionale* (national road). Still, also when three examples are provided, as in FSP2, the model answer, i.e., *nazione* (nation), presents a partial clue overlapping, that is it is derived by the first clue.

Another game resolved in ZSP1 and failed in the other settings is the game ID 152. In this case, the correct answer, i.e., *analisi* (analysis), is changed into (i) *algebra* (algebra), an hyponym of one of the [CLUES], that is *matematica* (mathematics) in ZSP2; (ii) *matematica*, that is one of the clues, when the model is provided with one example; (iii) *statistica* (statistics), another hyponym of mathematics, when we include three examples in the prompt.

Examples are proven to be useful for the correct solution in the game ID 47 for the FSP1 setting, while, in the ZSP results, the model provides a synonym of one of the clues for both settings, i.e., *celebrazione* (celebration) from the clue *festa* (party), and in the FSP2 setting the answer is *concorrenza* (competition), that does not seem having any relations to the clues.

GPT-4 GPT-4 presents some consistency between the correct results presented in ZSP1 and FSP2 in the game ID 37 and ID 135.

With reference to the use of the examples, it is worth noticing that in the game ID 59, GPT-3.5 solves the game when provided with one example and fails with three examples, GPT-4 needs three examples to give the correct [SOLUTION], while with one example the answer is *partito* (political party or left), that could be the result from a clue interference comning form the word *festa* (party), as in *festa di partito* (political convention).

Bard As already stated, Bard presents a high number of complete clue overlapping solutions, mainly in ZSP1. In some cases, the model is consistent in this error. For instance, the aforementioned ID 2 incorrect answer *saldo* (discount) is proposed in all settings, but in ZSP2, when the model proposes another clue as [SOLUTION], i.e., *fare* (to do). This type of error may indicate that the model does not understand the prompt.

This model presents consistency across the settings in ID 153. Indeed, Bard proposes the same correct [SOLUTION], i.e., *viso* (face), in all settings but ZSP1, when the output is *notte* (night), that

[18] It is worth stressing that solution including *agente* could belong also to an A-B pattern with the same meaning, as in *agente assicurativo*

seems completely out of context considering the provided [CLUES], that are *lavare* (to wash), *nuovo* (new), *espressione* (look), *maschera* (mask), *pallido* (pale).
Another case of consistency occurs in the game ID 204 in two settings, namely ZSP2 and FSP1, as the same correct answer *museo* (museum) is provided. In the remaining settings, the proposed outputs are *luna* (moon) in ZSP1, resulting from a clue intereference due to the presence of *notte* (night) among the [CLUES], and *uovo* (egg) in FSP2, that could be related to the clue *sale* (salt).

Gemini-Pro Gemini-Pro results show consistency, that is the provided [SOLUTION] is the same correct answer in three settings out of four, just in only one case that is ID 117.

- **ID 117**
  [CLUES]: *pesce* (fish), cary grant, *domestico* (domesticated), *donna* (woman), zorba
  [SOLUTION]: *gatto* (cat)

The model is consistent in all settings, but in ZSP1, when the proposed [SOLUTION] is *attore* (actor), due to the presence of the clue cary grant. The same incorrect answer is given also by GPT-3.5 in the same setting.
In another case, there is consistency among the results over the different settings. This is ID 162 in both FSP settings, but not in ZSP1 and ZSP2.

- **ID 162**
  [CLUES]: *finire* (to finish), *tutta* (entire), *brillante* (bright or comic) *italiana* (Italian), *maschera* (mask)
  [SOLUTION]: *commedia* (comedy)

Also Gemini presents consistency in the game ID 204, but, contrary to Bard results, in different settings. Indeed, this model provides the correct answer *museo* (*museum*) in ZSP1 and FSP2. While in the remaining settings, the proposed [SOLUTION] is *arte* (art) in ZSP2 and *mostra* (exhibition) in FSP1, both as results of a clue interference coming from the word moma.

6. Conclusion

In this paper, we present a series of experiments to investigate the reasoning skills and game-solving skills of four different LLMs (Bard, GPT-4, Gemini-Pro, and GPT-3.5) on a language game task called *GhigliottinAI*. We elicited solutions from the LLMs using different prompts in both zero-shot and few-shot settings. Specifically, for the few-shot setting, we provided both a game instance with its solution and three game instances with solutions from the GhigliottinAI task training set in the prompt.
As shown in Section 4.4, the performance achieved by the LLMs is quite low compared to the performance reported by other artificial players discussed in Section 3. In particular, the best performing artificial player (*Mago della Ghigliottina*) achieves an accuracy score of .68 compared to the two LLMs (GPT-4 and Gemini-Pro) that performed best in the FSP2 setting with an accuracy score of .022. Furthermore, in Section 5, we provide an analysis that aims to count the shared solutions proposed by the different LLMs to highlight how LLMs belonging to the same family have similar behaviors in solving the games.
As mentioned in Section 3 and Section 4.3, the game instances together with their solutions form a linguistic phenomenon known as MWE. This implies that in addition to evaluating reasoning abilities, the LLMs were also subjected to a test that assessed their knowledge of linguistic and statistical phenomena such as: word co-occurrence in frequent collocations or idioms, word similarity or word relatedness and semantic relationship of clues with solutions. In this context, in Section 5, we offered an analysis based on error types that can be explained by different levels of linguistic features.
While this type of analysis provides some preliminary insights into the results proposed by the LLMs, we plan to further investigate the behavior of LLMs in the GhigliottinAI task in the future. For example, to better evaluate the game-solving, reasoning abilities and linguistic phenomena knowledge of different LLMs, we plan to design prompts that elicit multiple solutions ranked by probability for each game instance, in order to rank the LLM proposals. In the process of eliciting diverse solutions ordered by probabilities, we also plan to design prompts with instructions that provide more linguistic context for the LLMs. Furthermore, since in this paper we only exploited two types of prompting techniques, we plan to refine the solution generation through Prompt Chain-of-Thought (Wei et al., 2022) and information retrieval from freely available corpora for the Italian language through Retrieval Augmented Generation (RAG) (Gao et al., 2023).

7. Acknowledgements

Authorship attribution is as follows: Raffaele Manna is author of Section 4 and Section 6, Maria Pia di Buono is author of Section 5, and Johanna Monti is author of Section 1, 2, and 3.
Maria Pia di Buono has been supported by Fondo FSE/REACT-EU - Progetti DM 1062 del 10/08/2021 "Ricercatori a Tempo Determinato di tipo A) (RTDA)". Azione IV.4 - Dottorati e contratti

di ricerca su tematiche dell'innovazione/Azione IV.6 - Contratti di ricerca su tematiche Green.

8. Bibliographical References

Pierpaolo Basile, Marco de Gemmis, Pasquale Lops, and Giovanni Semeraro. 2014. Solving a Complex Language Game by Using Knowledge-based Word Associations Discovery. *IEEE Transactions on Computational Intelligence and AI in Games*, 8(1):13–26.

Pierpaolo Basile, Marco De Gemmis, Lucia Siciliani, and Giovanni Semeraro. 2018. Overview of the Evalita 2018 Solving Language Games (NLP4Fun) Task. *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12:75.

Pierpaolo Basile, Marco Lovetere, Johanna Monti, Antonio Pascucci, Federico Sangati, and Lucia Siciliani. 2020. Ghigliottin-AI@ EVALITA2020: Evaluating Artificial Players for the Language Game “La Ghigliottina”. *EVALITA Evaluation of NLP and Speech Tools for Italian-December 17th, 2020*, page 345.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-Shot Learners. *Advances in neural information processing systems*, 33:1877–1901.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke Van Der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword Expression Processing: A Survey. *Computational Linguistics*, 43(4):837–892.

Nazareno De Francesco. 2020. GUL. LE. VER@ GhigliottinAI: A Glove based Artificial Player to Solve the Language Game “La Ghigliottina”. *EVALITA Evaluation of NLP and Speech Tools for Italian-December 17th, 2020*, page 356.

Marco Ernandes, Giovanni Angelini, and Marco Gori. 2008. A Web-based Agent Challenges Human Experts on Crosswords. *AI Magazine*, 29(1):77–77.

David Ferrucci, Anthony Levas, Sugato Bagchi, David Gondek, and Erik T Mueller. 2013. Watson: Beyond Jeopardy! *Artificial Intelligence*, 199:93–105.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint arXiv:2312.10997*.

Katikapalli Subramanyam Kalyan. 2023. A Survey of GPT-3 Family Large Language Models Including ChatGPT and GPT-4. *Natural Language Processing Journal*, page 100048.

K Lam, David M Pennock, Dan Cosley, Steve Lawrence, et al. 2012. 1 Billion Pages= 1 Million Dollars? Mining the Web to Play" Who Wants to be a Millionaire?". *arXiv preprint arXiv:1212.2477*.

Michael L Littman, Greg A Keim, and Noam Shazeer. 2002. A Probabilistic Approach to Solving Crossword Puzzles. *Artificial Intelligence*, 134(1-2):23–55.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? *arXiv preprint arXiv:2202.12837*.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large Language Models: A Survey. *arXiv preprint arXiv:2402.06196*.

Piero Molino, Pasquale Lops, Giovanni Semeraro, Marco de Gemmis, and Pierpaolo Basile. 2015. Playing with Knowledge: A Virtual Player for “Who Wants to Be a Millionaire?” that Leverages Question Answering Techniques. *Artificial Intelligence*, 222:157–181.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Basile Pierpaolo, Lovetere Marco, Johanna Monti, Antonio Pascucci, Sangati Federico, Siciliani Lucia, et al. 2020. Ghigliottin-AI@ EVALITA2020: Evaluating Artificial Players for the Language Game “La Ghigliottina”. In *CEUR WORKSHOP PROCEEDINGS*, pages 345–348. AILC-Associazione Italiana di Linguistica Computazionale.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Computational Linguistics and Intelligent Text Processing: Third International Conference, CICLing 2002 Mexico City, Mexico, February 17–23, 2002 Proceedings 3*, pages 1–15. Springer.

Federico Sangati, Antonio Pascucci, Johanna Monti, et al. 2018. Exploiting Multiword Expressions to Solve “La Ghigliottina”. In *Proceedings* *of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, volume 2263, pages 258–263. Accademia University Press.

Federico Sangati, Antonio Pascucci, Johanna Monti, et al. 2020. “Il Mago della Ghigliottina”@ Ghigliottin-AI: When Linguistics Meets Artificial Intelligence. In *CEUR WORKSHOP PROCEEDINGS*. AILC-Associazione Italiana di Linguistica Computazionale.

Giovanni Semeraro, Marco de Gemmis, Pasquale Lops, and Pierpaolo Basile. 2012. An Artificial Player for a Language Game. *IEEE Intelligent Systems*, 27(05):36–43.

Giovanni Semeraro, Pasquale Lops, Pierpaolo Basile, Marco De Gemmis, et al. 2009. On the Tip of My Thought: Playing the Guillotine Game. In *IJCAI*, pages 1543–1548.

Jianing Wang, Qiushi Sun, Nuo Chen, Xiang Li, and Ming Gao. 2023. Boosting Language Models Reasoning with Chain-of-Knowledge Prompting. *arXiv preprint arXiv:2306.06427*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Georgios N Yannakakis and Julian Togelius. 2018. *Artificial Intelligence and Games*, volume 2. Springer.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36.