

Analysis of Material Facts on Financial Assets: a Generative AI Approach

Gabriel Assis¹, Daniela Vianna⁴, Gisele L. Pappa³, Alexandre Plastino¹,
Wagner Meira Júnior³, Altigran Soares da Silva², Aline Paes¹

¹ Institute of Computing, Universidade Federal Fluminense, Niterói, RJ, Brazil

² Institute of Computing, Universidade Federal do Amazonas, Manaus, AM, Brazil

³ Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil

⁴ JusBrasil, Brazil

*assisgabriel@id.uff.br, daniela.vianna@jusbrasil.com.br, glpappa@dcc.ufmg.br,
plastino@ic.uff.br, alti@icomp.ufam.edu.br, meira@dcc.ufmg.br, alinepaes@ic.uff.br*

Abstract

Material facts (MF) are crucial and obligatory disclosures that can significantly influence asset values. Following their release, financial analysts embark on the meticulous and highly specialized task of crafting analyses to shed light on their impact on company assets, a challenge elevated by the daily amount of MFs released. Generative AI, with its demonstrated power of crafting coherent text, emerges as a promising solution to this task. However, while these analyses must incorporate the MF, they must also transcend it, enhancing it with vital background information, valuable and grounded recommendations, prospects, potential risks, and their underlying reasoning. In this paper, we approach this task as an instance of controllable text generation, aiming to ensure adherence to the MF and other pivotal attributes as control elements. We first explore language models' capacity to manage this task by embedding those elements into prompts and engaging popular chatbots. A bilingual proof of concept underscores both the potential and the challenges of applying generative AI techniques to this task.

Keywords: financial asset narratives, material fact, text generation, natural language instructions

1. Introduction

Material Facts (MF) are critical disclosures issued by companies and investment funds (Pallister and Law, 2006). They hold paramount importance in the market due to their potential to affect asset values. Owing to their substantial impact on the financial sector, these announcements are mandatory by law to maintain market transparency (Eastbrook and Fischel, 1984; Mahoney, 1995; Boatright, 2013). Their scope varies from alterations in a company's shareholder composition to significant acquisitions and disposals. Typical examples include executing agreements for transferring share control, changes in partnership, renegotiating debts, managing stock splits or mergers, reporting profits or losses, distributing dividends, adjusting financial forecasts, or initiating bankruptcy proceedings (Finnerty, 2013).

The mandatory announcement of material facts implicates a vast amount of documents. To have an idea, in the last trimester of 2023, the US Securities and Exchange Commission (SEC) had registered more than 23,000 material facts related to worldwide companies¹. After the announcements, financial analysts and other industry experts might produce an analysis relating the MF to financial assets. Arguably, the analysis is not only an expansion of the MF as they also might implicitly incorporate an-

alysts' background and other relevant events that are not explicitly related to the announcement (Subramanyam, 2014; Revsine et al., 2021).

This way, writing financial asset analyses from an MF presents various challenges. Material facts vary widely – some are brief, while others are lengthy and detailed; although all facts must be disclosed, their importance and impact can differ (Reichert and Perlin, 2022). Arguably, crafting accurate and valuable asset analyses from an MF requires additional information from other sources to point out how the MF might influence an asset, considering the market position, prospects, and potential risks (Chauvin and Hirschey, 1993). Additionally, the analysts' perspectives can bias the final recommendations, potentially compromising the accuracy of the information provided to the public (Hawes, 1975; Boatright, 2013). Furthermore, investment firms typically charge for access to their reports. While they are justified in doing so, this practice can limit the accessibility of financial information for those unable to afford their fees.

Conversely, with the advent and growing popularity of Generative AI powered by large language models (LLMs), automatic text generation (Erdem et al., 2022) has achieved remarkable results. However, generating a text that complies with another piece of information – as in financial asset analyses – still challenges modern LLMs (Peng et al., 2023; Zhang et al., 2024a). While the generated text must convey the material fact, it must, in essence,

¹<https://www.sec.gov/edgar/search/>

discuss its implications, bringing related information, including extended analyses and correlating important related events. Arguably, while LLMs are pre-trained with a massive amount of text and modern techniques offer the possibility of expanding them with Retrieval Augmented Generation (Lewis et al., 2020b) and reasoning abilities (Wei et al., 2022), they might still not be fully equipped to deal with the listed challenges.

This paper formalizes this task as an instance of controllable text generation (CTG) (Prabhumoye et al., 2020; Zhou et al., 2023; Zhang et al., 2024b). While previous works have explored several forms of CTG with LLMs (Prabhumoye et al., 2020; Zhang et al., 2024b; Zhou et al., 2023), most of the content-based category relies on simple elements, such as terms and topics, or self-contained texts (Liu et al., 2018; Yan, 2022). They have not examined such a diverse and, at the same time, untied element of control as a material fact.

Our initial strategy is implementing this task with crafted structured prompts embedded with a material fact and leveraging the in-context learning emerging abilities of state-of-the-art chatbots (Brown et al., 2020b; Dong et al., 2023) to write financial asset analyses. This preliminary investigation focuses on assessing the abilities of state-of-the-art LLMs to produce analyses solely based on their prior training stages.

Our proof of concept adopts a bilingual approach, generating analyses in both English and Portuguese, thereby probing the capabilities of LLMs in linguistic contexts beyond their primary training. To assess the efficacy of our method, we developed a proof of concept with a small set of reference financial analyses, sourced from reputed investment analyses. Then, we compare them against the outputs of various chatbots, encompassing a spectrum from open to closed systems and from small to large-scale models. We include two models of the GPT family (Ouyang et al., 2022; OpenAI, 2023), three Mistral models (Jiang et al., 2023, 2024), three Llama models (Touvron et al., 2023) and the recently released Gemini-Pro (Google, 2023).

While LLMs often blur the line between human and machine-generated texts, accurately assessing their quality remains challenging, especially in sensitive and specialized fields like finance. This way, this paper relies on traditional and modern text generation metrics to evaluate the output of chatbots in contrast to reference reports. This comparative analysis aims to assess the challenges and potential of the proposed task and at which point LLMs can tackle it.

To sum up, this paper contributes with

1. A novel task proposal for automatically generating asset analysis from material facts.
2. The task implementation leveraging generative AI, guided by well-crafted instruction incorporating a material fact.
3. A proof of concept with LLMs encompassing analyses in Portuguese and English.
4. We explore this problem with nine chatbots and evaluate the results with classical and state-of-the-art text generation metrics.

2. Related Work

2.1. Automating Financial Narratives

Prior research has investigated methods for generating financial reports from different inputs. One of the key areas of focus is generating reports from tabular data using table-to-text (TTG) techniques (Kale and Rastogi, 2020). Cm et al. (2023), for instance, explored pre-trained TTG transformers and how selecting crucial, correlated table attributes can aid in producing accurate money laundering reports. While material facts are typically presented in textual format, they may include numerical values both within tables and the text itself. Therefore, incorporating logical steps to integrate these values into reports accurately will be crucial in future efforts.

More aligned with our method, Yan (2022) developed a technique for creating financial reports from brief news articles. They focused on learning separate latent variables that capture the themes of the input news and the intended reports. The aim is to incorporate the natural uncertainty in reports, acknowledging that human experts contribute diverse perspectives and approaches to their analysis. Material facts may also consist of brief texts, but sometimes they have detailed information. Our proposed task targets producing reports that capture the source document's essential elements while enriching it with additional discussion and inferred insights. We begin our exploration of this issue by leveraging chatbots' innate capability to address these complexities.

2.2. Controllable Text Generation from Content

Frequently, text generation requires that the output agrees with a predefined specific element, such as style, structured data, or content (Erdem et al., 2022). Most recent work that generates controlled texts conditions the input to the required attribute (Prabhumoye et al., 2020). Zhang et al. (2024b) divides the strategies into the three following: (a.) adjusting (some of) pre-trained language models weights to produce texts with specific features (Ziegler et al., 2019; Liu et al., 2020), (b.) training controllable models with injecting controllers (Wang et al., 2021; He, 2021; Chan et al.,

2021), and (c.) post-processing PTLM signals that work only when decoding texts to incorporate the desired attributes (Hua and Wang, 2020; Dathathri et al., 2020). In our case, the report must be conditioned on the material fact while conveying related information. Our strategy aligns with the third aforementioned approach, as we include the material fact text in the prompt body and analyze whether chatbots attain it without further control.

3. Task Formulation

This paper introduces a novel task of generating analyses on financial assets (e.g., stocks, funds, private pensions, etc.) from material facts, leveraging text generation techniques. We formulate the task as an instance of controllable text generation, named *material fact controllable text generation* (MF-CTG). This way, MF-CTG is defined as $P(Y|MF, C, \mathbf{A})$, where the controllable elements are the material fact source text $MF = x_1 \dots x_m$, the company's name releasing it (C), and a set of other controllable attributes ($\mathbf{A} = \{A_1, A_2, \dots, A_k\}$), which could be style, impartial tone, structure, among others. The goal of the task is to generate a financial asset analysis report Y according to a vocabulary \mathcal{V} where $Y = y_1 \dots y_m$ and $y_k \in \mathcal{V}$.

This paper addresses this task using prompt-based generative pre-trained language models (PTLM). Therefore, \mathcal{V} is the PTLM vocabulary, and MF , C , and \mathbf{A} are included into a prompt PM , together with other elements, for example, context and instruction. This way, the task is $P(Y|PM) = \prod_{i=1}^n p(y_i|y_{<i}, PM)$, where $PM = \mathbf{z}_1 \dots \mathbf{A} \dots \mathbf{z}_1 \dots C \mathbf{z}_j \dots MF \mathbf{z}_w$ with \mathbf{z}_i being (possibly empty) sequences of words representing other information added to the prompt.

4. Instance and Evaluation of Material Fact Controllable Text Generation

This section describes our proposed method to address a concrete instance of MF-CTG and an evaluation routine. The method consists of the following procedures: (A.) Prompt Crafting, that assembles a prompt incorporating the related material fact, company's name, and analysis format as the controllable elements, among additional text;

(B.) Analyses Generation, rooted in activating the generative AI models through the prompt; and (C.) Analyses Evaluation, to evaluate the output analyses with automatic metrics. Those metrics assess analyses' lexical and syntactic aspects and rate semantic conformity regarding a reference report. While (A.) and (B.) implement MF-CTG, (C.) is responsible for evaluating its feasibility. Figure 1 depicts an overview of the proposed method.

4.1. Prompt Crafting

How a prompt is constructed significantly impacts the effectiveness of an LLM in performing downstream tasks (Liu et al., 2023a). In this context, prompt engineering techniques potentially optimize a model's performance. Those strategies encompass adding personas, using different delimiter symbols, incorporating reference materials, integrating examples for in-context learning, and outlining steps for task execution, often referred to as chain of thought (White et al., 2023; Brown et al., 2020a; Wei et al., 2022). While embracing those prompt engineering techniques, our strategy also entails directly integrating the controllable elements (material fact, its originating company's name and the format) into the prompt.

After conducting a series of preliminary assessments using these techniques, we observed that incorporating a persona and adding detailed instructions on the document's intended audience and format, along with an explicit formatting template, led to more appropriate responses. Specifically, the models produced content that was not only correctly formatted but also returned an analysis that considered the provided material fact and showed adherence to the vocabulary. Nonetheless, it is still crucial to thoroughly evaluate the content quality and the overall analysis. Further details on this are discussed in Section 4.3. The final prompt is as follows²:

```
### CONTEXT:
YOU ARE A FINANCIAL ANALYST WITH A BACKGROUND IN ECONOMICS WHO WRITES FOR A GENERAL INVESTOR AUDIENCE.

### INSTRUCTION:
WRITE AN ANALYSIS ABOUT THE COMPANY CONSIDERING THE MATERIAL FACT AND FOLLOW THE DETERMINED FORMAT. THE ANALYSIS MUST CONTAIN ALL THE ELEMENTS SPECIFIED IN THE FOLLOWING FORMAT.

### FORMAT:
<TITLE>
<BODY OF THE ANALYSIS>
<RECOMENDATION>

### COMPANY:
{COMPANY}

### MATERIAL FACT:
{SOURCE_DOCUMENT}

###RESPONSE:
```

4.2. Analyses Generation

Generating analyses from material facts with LLMs involves several factors. Those include the model's ability to process the given material fact and derive contextually pertinent information to compose a robust and relevant analysis. Therefore, assessing

²See Appendix B for the Portuguese prompt.

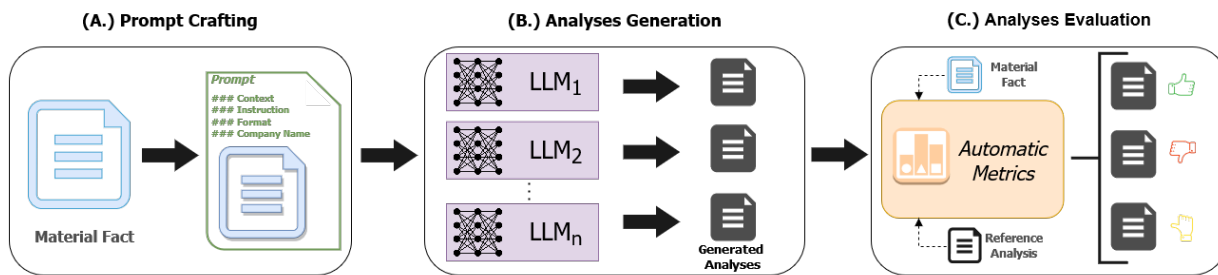


Figure 1: Schematic pipeline of the proposed MF-CTG solution.

the effectiveness of different models in carrying out the proposed MF-CTG task is critical to determine its feasibility.

The evaluation encompassed models based on the popular GPT architecture, specifically (i.) GPT-4 (OpenAI, 2023) and (ii.) GPT-3.5 (Ouyang et al., 2022). Additionally, the (iii.) Gemini-Pro 1.0 model (Google, 2023), known for its remarkable performance in recent benchmarks, and models from the Mistral family — (iv.) Mistral Medium, (v.) Mixtral 8x7B (Jiang et al., 2024), and (vi.) Mistral 7B (Jiang et al., 2023) — were included. We also adopted models from the Llama 2 series (Touvron et al., 2023), the (vii.) Llama 2 7B, (viii.) Llama 2 13B, and (ix.) Llama 2 70B. All models were utilized in their chat or instruction-based versions.

The selected models vary significantly in size, with their number of parameters spanning from a few billion to over a trillion. They also differ in terms of availability, including both closed (models i. to iv.) and open (subsequent models) sources. We aimed to assess how these variations affect the MF-CTG in both English and Portuguese. In this context, another aspect under investigation is the capability to generate analyses in a language for which all models were not predominantly pre-trained.

4.3. Analyses Evaluation

Evaluating texts produced by Generative AI methods remains a significant challenge to NLP (Kryscinski et al., 2019; Liu et al., 2023b). It is even more critical when considering texts from specific areas, such as finances. Nevertheless, characteristics addressed in NLG tasks, like summarization, transduction, and creative writing, establish a robust foundation for automated evaluation. Some critical criteria for evaluation encompass lexico-syntactic structure; fluency, which evaluates the clarity of the writing (Zhong et al., 2022); consistency, which checks how well the generated text aligns with the source document (Cao et al., 2020); coherence, assessing the logical connection of information (Ye et al., 2021); relevance, measuring the pertinence of the generated text to the critical infor-

mation based on the reference (Zhong et al., 2022); and groundedness, evaluating the text’s ability to reflect reasoned knowledge from the source document (Dinan et al., 2019). These criteria are crucial for the MF-CTG task, as they affect the report’s readability and can influence its credibility.

A widely adopted approach relies on metrics that measure the similarity of generated texts with references written by humans (Deng et al., 2021). In this context, our work incorporates the material fact and analyses gathered from financial firms’ websites, serving as reference reports.

We apply three groups of metrics. The first group regards semantic aspects with the following metrics: BERTScore (Zhang et al., 2020), which computes similarity based on contextualized embeddings; and BARTScore (Yuan et al., 2021), which proposes a unified evaluator based on the likelihood of the encoder-decoder model upon which it is based, and changes in the combination of its inputs. In our evaluation strategy, we propose segmenting both the evaluated text and the reference text. Thus, for an analysis report denoted as a , we define $\{a_t, a_o, a_b, a_c\}$ representing its title, overview, body, and conclusion, respectively. Similarly, for a reference denoted as r , we have $\{r_t, r_o, r_b, r_c\}$. We then calculate the scores for each pair $score(a_t, r_t)$, $score(a_o, r_o)$, $score(a_b, r_b)$, and $score(a_c, r_c)$, and subsequently calculate the average of the previous results. We conjecture that this method allows a more refined evaluation between the components of the analysis report, while also addressing the limitation of the context window often found in using these metrics with lengthy documents.

Conversely, UniEval (Zhong et al., 2022) was employed to evaluate fluency and coherence. This metric utilizes a binary question-answering (QA) pipeline built upon a generative LLM to calculate its value leveraging the probabilities of responses to questions like “*Is this text fluent?*”. Given its evaluation method and the aspects this metric covers are inherent to the text as a single piece, the previous segmentation approach was not applied.

The second group includes morphological and parsing analyses, including tokenization, part-of-

speech tagging, and dependency parsing-related metrics conveyed in UDPipe (Straka et al., 2016). It relies on the Universal Dependencies treebank annotations that include analyzers for both English and Portuguese. We compute the number of sentences, tokens, tokens per sentence, and the mean dependency distance (MDD). This last one aims to predict the syntactic difficulty of sentences according to psycholinguistics experiments (Liu, 2008).

Finally, the third group focuses only on the Portuguese analyses as we leverage the large set of metrics provided in NILC-metrix portal (Leal et al., 2023)³. Those metrics extract values from several linguistic proxies to assess morphosyntactic, cohesion, coherence, and textual complexity information. Once again, we compare the automatically generated analysis reports with the reference. We compute the metrics for the following groups: Referential Cohesion (seven metrics), Syntactic Complexity (27 metrics), Morphosyntactic Information (42 metrics), and Readability (five metrics).

5. Experimental Setup

This section describes the process of gathering material facts and references, the experimental settings employed in the inference process with LLMs, and further details on implementing the metrics.

5.1. Data

Collection Methodology Our evaluation includes documents in both English and Portuguese. To find openly accessible analyses in English, we benefit from Yahoo Finance, a popular tool in the financial context for indexing news and public reports⁴. We adopted keywords associated with typical topics in material facts as filters, e.g., reports of changes in partnership or organizational restructuring. More examples can be found in Section 1. Subsequently, the material facts issued on the same day or the day before by the entities mentioned in the reports were reviewed through the public system of the SEC⁵. The objective is to match the reports with the forms that contain this specific type of information, namely Forms 6-K for foreign companies, and Forms 8-K for US-based companies.

The reports in Portuguese were directly collected from the websites of financial analysis firms. The same keyword strategy was adopted. Moreover, the system of the regulatory agency equivalent to the SEC in Brazil, the *Comissão de Valores Mobiliários (CVM)*⁶, was used for the collection of the

material facts, in Portuguese, “*Fatos Relevantes*” forms.

Companies and Material Fact Selection We selected two reference analyses for each language. For Portuguese, the companies examined are BTG Pactual, a Brazilian investment bank specializing in investments and venture capital, and Eneva, a comprehensive Brazilian energy company engaged in power generation, oil and gas exploration and production, and electricity trading. For BTG Pactual, the critical event highlighted in the material fact was the acquisition of three properties, representing a multimillion-BRL transaction. In the case of Eneva, the significant event was a report on the rejection of a previously attempted merger by the company.

The cases in English concern Petrobras, Brazil’s largest oil company, a publicly traded corporation operating in the oil, natural gas, and energy sectors. The other company is Twilio, which offers communication tools and services through service APIs. In the case of Petrobras, the pertinent fact was the announcement of the intention to acquire the Jasper Block in the Campos Basin. Meanwhile, Twilio announced a layoff in its global workforce and the integration of Twilio Flex.

This selection requires that the models show a broad range of abilities and knowledge. They must not only consider the events and companies involved but also reason about factors such as geographical nuances, and the sizes and sectors of the companies. The MFs are presented entirely in the Appendix C.

5.2. Generative Models Inference

We conduct inference on generative models through ChatBot Arena⁷ (Zheng et al., 2023), an LLM benchmark platform that features comparisons between models in a crowdsourced manner. The platform provides access to models such as *gpt-4-1106*, *gpt-3.5-turbo-0613*, *gemini-pro-dev-api*, *mistral-medium*, *mistral-8x7b-instruct-v0.1*, *mistral-7b-instruct*, *llama-2-7b-chat*, *llama-2-13b-chat*, and *llama-2-70b-chat*, all of which were adopted in this work. Our choice is based on the high computational costs of running huge models. It also fits the objective of conducting a preliminary evaluation of the models’ ability to generate financial analyses.

As well known, the choice of hyperparameters such as *temperature* and *top_p* significantly influences the responses generated by models (Döderlein et al., 2022). To balance between aspects like creative writing and truthfulness, these parameters were set at $\{temperature = 0.3, top_p = 0.4\}$. The

³<http://fw.nilc.icmc.usp.br:23380/nilcmetrix>

⁴<https://finance.yahoo.com/>

⁵<https://www.sec.gov/edgar/>

⁶<https://sistemas.cvm.gov.br/>

⁷<https://arena.lmsys.org/>

max_tokens parameter was set to 2000, the highest possible value in Arena, to not limit the models in terms of the length of the generated analyses.

5.3. Metrics Implementation

Concerning the implementation of automatic semantic metrics, we leveraged models that we conjecture to have a more suited semantic representation for metrics that measure similarity, given the financial context, whether through the vector representation of contextualized embeddings or the likelihood of tokens. Consequently, we employed the BERTScore metric on FinBERT (Araci, 2019), a model pre-trained on a financial *corpus*. However, the vocabulary of this model is limited to English. Thus, for analyses in Portuguese, we used the multilingual version of BERT (Devlin et al., 2019)⁸, given that the metric’s implementation does not support a specific model for Portuguese. For BARTScore, which was originally proposed based on the encoder-decoder model BART (Lewis et al., 2020a), the implementation for English analysis was carried out on its version tuned on CNN and DailyMail news *corpus*. This choice was made due to the higher correlation with human evaluation reported by the original work (Yuan et al., 2021). For Portuguese, similarly to the approach taken for the previous metric, the multilingual version of the original model was employed⁹.

Additionally, the UniEval metric is originally proposed on a fine-tuned version of the T5 model (Rafael et al., 2020). Considering the limitation that may be associated with the number of tokens in lengthy financial analyses and to prevent information loss due to truncation, we implemented the original formulation proposed by (Zhong et al., 2022) within the probability results obtained by querying GPT-4 via its API. Further details in the Appendix D. Finally, the implementations of the other metrics and the other hyperparameters settings for all metrics were kept at their default values.

6. Proof of Concept Results

6.1. Quantitative Analysis

Table 1 showcases the results of the first group of metrics, assessing semantic aspects compared to reference reports. UniEval could not discriminate among the LLMs, as the probabilities returned by GPT-4 are always very close to 1. Conversely, BERTScore and BARTScore failed to conclusively identify a superior model, as scores were generally

⁸<https://huggingface.co/google-bert/bert-base-multilingual-cased>

⁹<https://huggingface.co/facebook/bart-large-cnn>

close and varied models excelled in different scenarios. However, two key insights emerged from the results: (i.) the values point out that the reports generated by LLMs closely resemble the reference collection, underscoring the potential of our proposed task, and (ii.) surprisingly, the metric scores for the Portuguese results are generally higher than those for English. This is notable considering the models applied for Portuguese are multilingual and were not explicitly optimized for the financial sector, unlike FinBERT.

Table 2 displays token and sentence count, average tokens per sentence, and the average dependency distance, together with the difference regarding the reference, for both Portuguese and English scenarios. In most instances, the number of tokens and sentences was lower than the reference. While this could be interpreted as greater conciseness, some significant differences indicate the possibility that the models might not have generated additional and relevant information beyond the content of the material fact. Conversely, the values for MDD are significantly close to the reference, which reinforces the notable ability of LLMs in generating texts with grammatical complexity similar to human-written texts (Liu et al., 2023b).

Finally, Table 3 brings the Euclidean distance between the reference and generated reports using the average of four groups of metrics from NILC-matrix, only for the Portuguese cases. Before computing the distance, we normalized the values between 0 and 1. Overall, the models performed closely to the reference, achieving small distances for most metric groups. The poorest performance was observed in referential cohesion for Eneva, where the distance reached half of its maximum potential value. Examining morphosyntax reveals an interesting case with BTG, in which the entire Llama series achieved below-average results, including the worst overall performance. This is evident from the analyses, which include examples such as “*um área*”, a gender mismatch in Portuguese. Consequently, writing in a language for which the models were not primarily pre-trained may pose a significant hurdle. The Mistral and GPT families each claimed half of the top spots. However, while a Mistral model was among the worst performers alongside Llama and Gemini, no GPT model fell into this category. Nonetheless, the open models show promise, delivering solid performances and allowing for further exploration due to their greater accessibility than GPT.

These metrics also reveal that size does not necessarily equate to consistent behavior. For instance, the 7B versions of Mistral and Llama exhibit several distant values in the BTG reports. Moreover, size is not always a determinant of performance, as pointed out by the results within the Llama family:

	Portuguese						English					
	BERTScore		BARTScore		UniEval		BERTScore		BARTScore		UniEval	
	BTG	Ene.	BTG	Ene.	BTG	Ene.	Petro.	Twillio	Petro.	Twillio	Petro.	Twillio
GPT-4-turbo	0.715	0.681	-7.295	-5.973	1.000	0.999	0.627	0.599	-3.188	-3.386	0.998	0.999
GPT-3.5-turbo	0.531	0.504	-6.524	-6.793	1.000	0.998	0.591	0.593	-3.633	-3.258	1.000	1.000
Gemini Pro	0.698	0.668	-7.675	-8.356	1.000	0.999	0.579	0.596	-3.556	-3.322	1.000	1.000
Mistral Medium	0.710	0.682	-6.474	-6.062	1.000	1.000	0.625	0.579	-3.164	-3.391	0.999	1.000
Mixtral 8x7B	0.710	0.667	-6.082	-6.546	1.000	1.000	0.627	0.610	-3.138	-3.257	1.000	1.000
Mistral 7B	0.710	0.672	-5.213	-6.023	0.998	0.997	0.635	0.599	-3.147	-3.252	1.000	1.000
Llama2 7B	0.697	0.675	-6.005	-6.118	1.000	1.000	0.575	0.599	-3.690	-3.300	1.000	1.000
Llama2 13B	0.696	0.682	-5.768	-5.803	0.999	0.999	0.589	0.615	-3.747	-3.294	1.000	1.000
Llama2 70B	0.715	0.667	-7.326	-5.566	0.993	0.990	0.581	0.600	-3.683	-3.247	1.000	1.000

Table 1: Comparative results of semantic evaluation metrics for Portuguese and English cases. It presents the recall values for BERTScore and BARTScore, alongside the UniEval average scores. All metrics except BARTScore are scored in the range of 0 to 1, with higher scores indicating better quality. Best-performing results are highlighted in **bold** for each case.

	Portuguese								English							
	#sent. (reference Δ)		#tokens (reference Δ)		#tokens/sent. (reference Δ)		MDD (reference Δ)		#sent. (reference Δ)		#tokens (reference Δ)		#tokens/sent. (reference Δ)		MDD (reference Δ)	
	BTG	Ene.	BTG	Ene.	BTG	Ene.	BTG	Ene.	Petro.	Twillio	Petro.	Twillio	Petro.	Twillio	Petro.	Twillio
GPT-4 turbo	18 (-22)	26 (-2)	356 (-254)	403 (-47)	19.778 (+4.528)	15.500 (-0.571)	3.827 (+0.280)	3.444 (-0.150)	26 (-38)	36 (-4)	437 (-279)	465 (+7)	16.808 (+5.620)	12.917 (+1.467)	4.132 (+0.527)	3.808 (+0.161)
GPT-3.5 turbo	24 (-16)	24 (-4)	341 (-269)	329 (-121)	14.208 (-1.042)	13.708 (-2.363)	3.796 (+0.249)	3.621 (+0.027)	30 (-34)	36 (-4)	438 (-278)	427 (+3.413)	14.600 (+3.413)	11.861 (+5.017)	3.827 (+0.222)	3.753 (+0.097)
Gemini Pro	18 (-22)	22 (-6)	235 (-375)	267 (-183)	13.056 (-2.194)	12.136 (-3.935)	3.531 (-0.016)	3.180 (-0.414)	22 (-42)	22 (-18)	343 (-373)	314 (+4.403)	15.591 (+2.823)	14.273 (+2.823)	3.892 (+0.286)	3.812 (+0.166)
Mistral Medium	20 (-20)	22 (-6)	320 (-290)	347 (-103)	16.000 (+0.750)	15.773 (-0.289)	3.410 (-0.138)	3.521 (-0.073)	34 (-30)	34 (-6)	445 (-271)	438 (+1.901)	13.088 (+1.432)	12.882 (+0.098)	3.703 (+0.098)	3.786 (+0.140)
Mixtral 8x7B	24 (-16)	26 (-2)	399 (-211)	274 (-176)	16.625 (+1.375)	10.538 (-5.533)	3.519 (-0.028)	3.383 (-0.211)	20 (-44)	30 (-10)	368 (-348)	384 (+7.213)	18.400 (+1.350)	12.800 (+0.518)	4.123 (+0.292)	3.938 (+0.292)
Mistral 7B	40 (0)	40 (+12)	615 (+5)	485 (+35)	15.375 (+0.125)	12.125 (-3.946)	3.763 (+0.216)	3.138 (-0.456)	32 (-12)	28 (-21)	505 (-51)	407 (+4.594)	15.781 (+3.086)	14.536 (+0.470)	4.076 (+0.308)	3.954 (+0.308)
Llama2 7B	28 (-12)	30 (+2)	391 (-219)	327 (-123)	13.964 (-1.286)	10.900 (-5.171)	3.422 (-0.126)	3.266 (-0.328)	20 (-44)	30 (-10)	379 (-337)	494 (+36)	18.950 (+7.763)	16.467 (+5.017)	3.821 (+0.216)	3.743 (+0.097)
Llama2 13B	26 (-14)	26 (-2)	377 (-233)	339 (-111)	14.500 (-0.750)	13.038 (-3.033)	3.495 (-0.053)	3.390 (-0.204)	32 (-32)	44 (+4)	469 (-247)	640 (+182)	14.656 (+3.704)	14.545 (+3.095)	3.704 (+0.099)	3.649 (+0.003)
Llama2 70B	32 (-8)	46 (+18)	418 (-192)	497 (+47)	13.063 (-2.188)	10.804 (-5.267)	3.468 (-0.080)	3.403 (-0.191)	36 (-28)	28 (-12)	445 (-271)	423 (+35)	12.361 (+1.174)	15.107 (+3.657)	3.663 (+0.058)	3.902 (+0.256)

Table 2: Comparative results of Morphological and parsing metrics computed with UDPipe.

the larger model does not necessarily outperform its smaller counterparts.

6.2. Manual Inspection of an Analysis

We selected the EN Petrobras analysis to take a fine-grained look at how a particular analysis addressed the material fact. One representative of each family was selected according to their performance on similarity and morphosyntactic results (Tables 1 and 2). Although Gemini did not achieve the best results, we also bring it here for comparison. The Appendix E discusses an AI-generated analysis and the reference. Regarding the format, all the analyses include a title, body, and recommendation followed or not by additional conclusions or (an attempt of) reasoning strategy.

Table 4 exhibits the titles of the material fact, the reference, and the four selected analyses. We notice that all models extensively add words and phrases from the MF title, whereas the reference title summarizes the main point more concisely. All the models include the company’s name and the expression “Production Sharing regime” directly from the MF title. The title with more words in common with the MF is Gemini, indicating a lower level of creativity in elaborating beyond merely reflecting

the title. Llama is almost the same, but at least it included an expression to suggest further discussion (*A Promising Move?*). Conversely, Mistral 7B included the block name (*Jaspe Block*) and GPT-4-Turbo also included the block location (*Campos Basin*), both of them mentioned in the MF body.

In terms of the main body of the analyses, while all the generated analyses address the primary subject of the material fact, they tend to be quite superficial and merely outline the anticipated outcomes of the acquisition. Moreover, they missed listing positive outcomes. For instance, none mentioned the potential for job creation or the advancement of technology when exploring the block, as the reference did.

While the reference report describes the block as “a geological treasure trove responsible for roughly 80% of Brazil’s oil output...” the machine-generated analysis lack more information about it. They only superficially mention the Jaspe Block as promising and the Campos Basin as prolific. The way we activate the LLMs does not provide them with direct access to such detailed information, even though it might be within their pre-training data. This gap suggests that the decoded analysis could benefit from external sources of information.

	Referential Cohesion		Syntactic Complexity		Morphosyntactic		Readability	
	BTG	Ene.	BTG	Ene.	BTG	Ene.	BTG	Ene.
<i>GPT-4-turbo</i>	0.3063	1.2193	0.0456	0.1103	0.2337	0.2474	0.0244	0.0216
<i>GPT-3.5-turbo</i>	0.1591	1.2652	0.0600	0.1044	0.3089	0.2930	0.0112	0.0138
<i>Gemini Pro</i>	0.2970	1.2110	0.0678	<u>0.1561</u>	0.2952	<u>0.4814</u>	<u>0.0465</u>	0.0373
<i>Mistral Medium</i>	0.1784	1.0560	0.0521	0.0869	0.3149	0.3437	0.0053	0.0011
<i>Mixtral 8x7B</i>	0.3716	1.2967	0.0758	0.1087	0.3022	0.3310	0.0263	0.0299
<i>Mistral 7B</i>	0.6112	1.2686	0.0577	0.0841	0.3386	0.3918	0.0244	<u>0.0381</u>
<i>Llama2 7B</i>	0.2789	<u>1.3034</u>	0.0956	0.1213	0.6108	0.3538	0.0061	0.0274
<i>Llama2 13B</i>	0.1904	1.2742	0.0683	0.0905	0.4670	0.3002	0.0293	0.0105
<i>Llama2 70B</i>	0.5066	1.1143	0.0624	0.1101	0.4207	0.4329	0.0204	0.0158
Mean	0.3222	1.2232	0.0650	0.1080	0.3658	0.3528	0.0215	0.0217
(std)	(0.1530)	(0.0853)	(0.0146)	(0.0219)	(0.1153)	(0.0729)	(0.0129)	(0.0125)

Table 3: Euclidean Distance between NILC-matrix results for BTG and Eneva automatically generated analyses and the reference analysis. The values in **bold** are the less distant to the reference, and the underline values are the more distant.

Source	Title
<i>Material Fact</i>	Petrobras expresses interest in area under the Production Sharing regime
<i>Reference</i>	Petrobras (PBR) Plans to Acquire Jaspe Block in Offshore Brazil
<i>Mistral 7B</i>	Petrobras' Interest in Jaspe Block Under Production Sharing Regime: An Analysis
<i>GPT-4-turbo</i>	Petrobras Eyes Expansion in Campos Basin with Interest in Jaspe Block under Production Sharing Regime
<i>Llama2 70B</i>	Petrobras Expresses Interest in Production Sharing Regime: A Promising Move?
<i>Gemini Pro</i>	Petrobras Expresses Interest in New Production Sharing Regime Block

Table 4: Titles of the Material Fact, reference analysis, and four selected models.

Lastly, some of the generated analyses mention risks and potential negative outcomes, albeit in a broad manner. For example, GPT-4 points out that “*exploration and development of new blocks carry inherent risks, including geological, technological, and market-related uncertainties*”. It also raises concerns about the global move towards renewable energy and the need to meet ESG standards. Mistral highlights “*competition from other oil and gas companies in the region and potential regulatory challenges*”. Yet, none address the need for the company to take care of and avoid negative environmental impacts, which is a possible outcome that the reference analysis emphasized. Llama and Gemini fail to cover challenges, with Gemini notably adopting a more positive tone, in line with the material fact.

These aspects underscore the challenges of the task we propose in this paper. The analysis should naturally align with the material fact, being compliant, consistent, and coherent. If one targets the reference analyses style, refining the models with examples, explicitly fine-tuning their weights, or enhancing the prompts with demonstrations is recommended.

7. Conclusions

This paper introduced a novel financial task: automatically generating financial asset analyses based on material facts. The task is approached as an

instance of controllable text generation, with the material fact (MF) and the company’s name serving as primary control elements, alongside other attributes like report structure and tone. We employed generative AI techniques, incorporating these control elements into the prompts. A bilingual proof of concept with four references, nine LLMs, and using semantic, morphological, and syntactic metrics, highlights the proposal’s potential and challenges. Among the models we tested, we highly recommend further exploration of Mistral 7B due to its impressive performance in both languages, coupled with the fact that it is openly available and free to use. As expected, GPT-4 also performed remarkably in the Portuguese analyses. Future directions include improving the analyzed information by incorporating relevant facts, gathering more data to make it possible to fine-tune the models, and investigating possible hallucinations. Although we have not discussed that in the paper, we noticed that one of the reports in Portuguese included unreal affirmations. We also plan to design more precise evaluation metrics tailored to the financial sector and achieve more fine-grained control over the generated text by tuning the models with more precise instructions.

Acknowledgments

This research was financed by the Brazilian research agencies CNPq, CAPES, and FAPERJ.

8. Bibliographical References

- Dogu Araci. 2019. [FinBERT: Financial Sentiment Analysis with Pre-trained Language Models](#). *CoRR*, abs/1908.10063.
- John R Boatright. 2013. *Ethics in finance*. John Wiley & Sons.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. [Language Models are Few-Shot Learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. [Factual Error Correction for Abstractive Summarization Models](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258, Online. Association for Computational Linguistics.
- Alvin Chan, Yew-Soon Ong, Bill Pung, Aston Zhang, and Jie Fu. 2021. [CoCon: A Self-Supervised Approach for Controlled Text Generation](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Keith W Chauvin and Mark Hirschey. 1993. Advertising, r&d expenditures and the market value of the firm. *Financial management*, pages 128–140.
- Harika Cm, Debasmita Das, Ram Ganesh V, Rakesh Kumar Ranjan, and Siddhartha Asthana. 2023. [LoKI: Money Laundering Report Generation via Logical Table-to-Text using Meta Learning](#). In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 104–110, Macao. -.
- Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. [Plug and Play Language Models: A Simple Approach to Controlled Text generation](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric Xing, and Zhiting Hu. 2021. [Compression, Transduction, and Creation: A Unified Framework for Evaluating Natural Language Generation](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7580–7605, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. [Wizard of Wikipedia: Knowledge-Powered Conversational Agents](#). In *International Conference on Learning Representations*.
- Jean-Baptiste Döderlein, Mathieu Acher, Djamel Eddine Khelladi, and Benoît Combe-male. 2022. [Piloting Copilot and Codex: Hot Temperature, Cold Prompts, or Black Magic?](#) *CoRR*, abs/2210.14699.
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. [A Survey for In-context Learning](#). *CoRR*, abs/2301.00234.
- Frank H Easterbrook and Daniel R Fischel. 1984. Mandatory disclosure and the protection of investors. *Virginia Law Review*, pages 669–715.

- Erkut Erdem, Menekse Kuyu, Semih Yagcioglu, Anette Frank, Letitia Parcalabescu, Barbara Plank, Andrii Babii, Oleksii Turuta, Aykut Erdem, Iacer Calixto, Elena Lloret, Elena Simona Apostol, Ciprian-Octavian Truica, Branislava Sandrih, Sanda Martincic-Ipsic, Gábor Berend, Albert Gatt, and Grazina Korvel. 2022. [Neural Natural Language Generation: A Survey on Multilinguality, multimodality, controllability and learning](#). *J. Artif. Intell. Res.*, 73:1131–1207.
- John D Finnerty. 2013. *Project financing: Asset-based financial engineering*, volume 852. John Wiley & Sons.
- Team Google. 2023. [Gemini: A Family of Highly Capable Multimodal Models](#). *CoRR*, abs/2312.11805.
- Douglas W Hawes. 1975. Truth in Financial Statements: An Introduction. *Vand. L. Rev.*, 28:1.
- Xingwei He. 2021. [Parallel Refinements for Lexically Constrained Text Generation with BART](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8653–8666, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Xinyu Hua and Lu Wang. 2020. [PAIR: Planning and Iterative Refinement in Pre-trained Transformers for Long Text Generation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 781–793, Online. Association for Computational Linguistics.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7B](#). *CoRR*, abs/2310.06825.
- Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L  lio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Th  ophile Gervet, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2024. [Mixtral of Experts](#). *CoRR*, abs/2401.04088.
- Mihir Kale and Abhinav Rastogi. 2020. [Text-to-Text Pre-Training for Data-to-Text Tasks](#). In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 97–102, Dublin, Ireland. Association for Computational Linguistics.
- Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. [Neural Text Summarization: A Critical Evaluation](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China. Association for Computational Linguistics.
- Sidney Evaldo Leal, Magali Sanches Duran, Carolina Evaristo Scarton, Nathan Siegle Hartmann, and Sandra Maria Alu  sio. 2023. [NILC-Matrix: assessing the complexity of written and spoken language in Brazilian Portuguese](#). *Language Resources and Evaluation*, pages 1–38.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. [BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich K  ttler, Mike Lewis, Wen-tau Yih, Tim Rockt  schel, Sebastian Riedel, and Douwe Kiela. 2020b. [Retrieval-augmented generation for knowledge-intensive NLP tasks](#). In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS’20*, Red Hook, NY, USA. Curran Associates Inc.
- Haitao Liu. 2008. [Dependency Distance as a Metric of Language Comprehension Difficulty](#). *The Journal of Cognitive Science*, 9:159–191.
- Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023a. [Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing](#). *ACM Comput. Surv.*, 55(9).
- Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. [Generating Wikipedia by Summarizing Long Sequences](#). In *6th International Conference on Learning Representations*,

- ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net.
- Ruibao Liu, Guangxuan Xu, Chenyan Jia, Weicheng Ma, Lili Wang, and Soroush Vosoughi. 2020. [Data Boost: Text Data Augmentation Through Reinforcement Learning Guided Conditional Generation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9031–9041, Online. Association for Computational Linguistics.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. [G-eval: NLG Evaluation using GPT-4 with Better Human Alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Paul G Mahoney. 1995. Mandatory disclosure as a solution to agency problems. *The University of Chicago Law Review*, 62(3):1047–1112.
- Team OpenAI. 2023. [GPT-4 Technical Report](#). *CoRR*, abs/2303.08774.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.
- John Pallister and Jonathan Law. 2006. A dictionary of business and management.
- Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.
- Shrimai Prabhumoye, Alan W Black, and Ruslan Salakhutdinov. 2020. [Exploring Controllable Text Generation Techniques](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1–14, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *The Journal of Machine Learning Research*, 21(1).
- Marcos Henrique Reichert and Marcelo Scherer Perlin. 2022. [What drives the release of material facts for brazilian stocks?](#) *Brazilian Review of Finance*.
- Lawrence Revsine, Daniel W Collins, and W Bruce Johnson. 2021. *Financial reporting & analysis*. McGraw-Hill.
- Milan Straka, Jan Hajič, and Jana Straková. 2016. [UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).
- KR Subramanyam. 2014. *Financial statement analysis*. McGraw-Hill.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open Foundation and Fine-Tuned Chat Models](#). *CoRR*, abs/2307.09288.
- Yufei Wang, Ian D. Wood, Stephen Wan, Mark Dras, and Mark Johnson. 2021. [Mention Flags \(MF\): Constraining Transformer-based Text Generators](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event*,

- August 1-6, 2021, pages 103–113. Association for Computational Linguistics.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. [Chain-of-Thought Prompting Elicits Reasoning in Large Language Models](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf El-nashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. [A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT](#). *CoRR*, abs/2302.11382.
- Sixing Yan. 2022. [Disentangled Variational Topic Inference for Topic-Accurate Financial Report Generation](#). In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 18–24, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Zheng Ye, Liucun Lu, Lishan Huang, Liang Lin, and Xiaodan Liang. 2021. [Towards Quantifiable Dialogue Coherence Evaluation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2718–2729, Online. Association for Computational Linguistics.
- Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. [BARTScore: Evaluating Generated Text as Text Generation](#). In *Advances in Neural Information Processing Systems*.
- Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2024a. [A survey of controllable text generation using transformer-based pre-trained language models](#). *ACM Comput. Surv.*, 56(3):64:1–64:37.
- Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2024b. [A Survey of Controllable Text Generation Using Transformer-based Pre-trained Language Models](#). *ACM Comput. Surv.*, 56(3):64:1–64:37.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [BERTScore: Evaluating Text Generation with BERT](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena](#). In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. [Towards a Unified Multi-Dimensional Evaluator for Text Generation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Wangchunshu Zhou, Yuchen Eleanor Jiang, Ethan Wilcox, Ryan Cotterell, and Mrinmaya Sachan. 2023. [Controlled Text Generation with Natural Language Instructions](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 42602–42613. PMLR.
- Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. 2019. [Fine-Tuning Language Models from Human Preferences](#). *CoRR*, abs/1909.08593.

A. Ethical Statement

The *corpora* that generative models were pre-trained on may harbor socio-economic partialities, which can lead to these biases being perpetuated in the financial analyses they produce. Additionally, as a task of creative writing, the output can sometimes be marred by the issue of hallucination, leading to the propagation of illusory and occasionally distorted information. We reinforce the recent literature that argues the need for thorough investigation in this aspect. Moreover, financial asset analyses significantly influence investors’ decision-making process, from the experienced to the beginners, which can lead to both monetary gains and losses. In this sense, we believe that the AI-generated analyses could play a new and crucial role by providing open and free content on which people can base their investment strategies. However, we note that recommendations from AI should not be accepted uncritically and emphasize the importance of consulting multiple sources of information.

B. Prompt in Portuguese

CONTEXTO:

VOCÊ É UM ANALISTA FINANCEIRO COM FORMAÇÃO EM ECONOMIA QUE ESCREVE PARA UM PÚBLICO DE INVESTIDORES GERAL.

INSTRUÇÃO:

ESCREVA EM PORTUGUÊS UMA ANÁLISE SOBRE A EMPRESA CONSIDERANDO O FATO RELEVANTE E SIGA O FORMATO DETERMINADO. A ANÁLISE DEVE CONTER TODOS OS ELEMENTOS ESPECIFICADOS NO SEGUINTE FORMATO.

FORMATO:

<TÍTULO>

<SENTENÇA QUE RESUMA A ANÁLISE>

<CORPO DA ANÁLISE>

<RECOMENDAÇÃO>

EMPRESA:

{COMPANY}

FATO RELEVANTE:

{SOURCE_DOCUMENT}

###RESPOSTA:

C. The Material Facts

C.1. BTG Pactual

A BTG PACTUAL SERVIÇOS FINANCEIROS S.A. DTVM, inscrita no CNPJ sob o nº 59.281.253/0001-23 (“Administradora”), e a BTG PACTUAL GESTORA DE RECURSOS LTDA., inscrita no CNPJ sob o nº 09.631.542/0001-37 (“Gestora”), na qualidade de administradora e gestora, respectivamente, do BTG PACTUAL LOGÍSTICA FUNDO DE INVESTIMENTO IMOBILIÁRIO inscrito no CNPJ sob o nº 11.839.593/0001-09 (“Fundo”), serve-se do presente para informar aos cotistas e ao mercado em geral quanto o seguem que: I. O Fundo celebrou, nesta data, Instrumentos Particulares (“Instrumento”), sujeitos a determinadas regras de confidencialidade, tendo por objeto a promessa de venda e compra de três imóveis localizados em São Paulo, sendo 70% em um raio de 30km, com área total de, aproximadamente, 233.000m² totalmente locadas (“Imóveis Performados”) e um projeto aprovado de, aproximadamente, 74.000 m² (em conjunto com os Imóveis Performados, “Imóveis”), pelo montante total de R\$ 760.000.000,00 (setecentos e sessenta milhões de reais) (“Preço da Venda”), a serem desembolsados da seguinte forma: (i) Primeira Parcela: R\$ 440.000.000,00 (quatrocentos e quarente milhões de reais) na data de fechamento; e (ii) Segunda parcela: R\$ 320.000.000,00 (trezentos e vinte milhões de reais) após 18 meses da data de fechamento, corrigidos pelo Índice Nacional de Preços ao Consumidor

Amplio (“IPCA”). II. O fechamento da operação está condicionado a verificação de condições precedentes usuais para este tipo de operação que, quando verificadas, serão comunicadas ao mercado. III. O pagamento parcelado atrelado ao recebimento total das receitas a partir da data de fechamento proporcionará ao Fundo um Yield estimado de 15% até o pagamento da parcela final. A receita estimada desta operação é de R\$ 0,19 por cota. IV. O cap rate envolvido na operação, ou seja, o valor de receita vigente sobre o Preço dos Imóveis Performados é de 9,2%. V. Por fim, a Gestora ressalta que a nova aquisição é resultado do trabalho ativo que vem realizando com o intuito de gerar valor para o Fundo e seus cotistas.

C.2. Eneva

Proposta Não-Vinculante para Fusão de Iguais com Vibra Energia S.A. – Resposta do Conselho de Administração da Vibra

Rio de Janeiro, 28 de novembro de 2023 – ENEVA S.A. (“Eneva” ou “Companhia”) (B3: ENEV3), em atendimento ao disposto no artigo 157, § 4.º da Lei n.º 6.404, de 15 de dezembro de 1976, e na Resolução CVM n.º 44, de 23 de agosto de 2021, e em continuidade ao fato relevante divulgado em 26 de novembro de 2023 a respeito do envio de proposta de combinação de negócios ao Conselho de Administração da Vibra Energia S.A. (“Proposta” e “Vibra”) (B3: VBBR3), vem comunicar a seus acionistas e ao mercado em geral que, por meio de correspondência recebida na data de hoje (cuja cópia consta anexa), a Vibra informou à Eneva que seu Conselho de Administração rejeitou a Proposta. A administração da Companhia avaliará tal resposta oportunamente e a Eneva se compromete a manter seus acionistas e o mercado em geral informados a respeito de novos desdobramentos relevantes a respeito deste tema na forma da lei e da regulamentação da Comissão de Valores Mobiliários – CVM.

C.3. Petrobras

Petrobras expresses interest in area under the Production Sharing regime

Rio de Janeiro, January 24, 2024 – Petróleo Brasileiro S.A. – Petrobras informs that it expressed today to the National Energy Policy Council (CNPE) its interest in the right of first refusal in a block to be tendered in the Permanent Offer System, under the Production Sharing Regime, under the terms of Law 12,351/2010 and Federal Decree 9,041/2017. Petrobras approved the expression of interest in the right of first refusal in the Jaspe block, located in the Campos Basin, considering the parameters

disclosed in CNPE Resolution No. 11, of December 20, 2023, published on December 27, 2023. The expression of interest is in line with the E&P strategy set out in SP 24-28+, focusing on profitable assets and replenishing oil and gas reserves. Material facts on the subject will be disclosed to the market in due course.

C.4. Twilio

Costs Associated with Exit or Disposal Activities.

On December 4, 2023, Twilio Inc. (the “Company”, “we” or “our”) committed to a further workforce restructuring plan (the “December Plan”) intended to streamline operations and accelerate the Company’s path to delivering profitable growth. The December Plan includes the elimination of approximately 5As as a result of the December Plan, the Company estimates that it will incur approximately 25–35 million in charges in connection with the workforce reduction, consisting of expenditures for employee transition, notice period and severance payments, employee benefits, and related facilitation costs, substantially all of which are expected to result in future cash outlays. The Company expects that the majority of the restructuring charges related to the December Plan will be incurred in the fourth quarter of 2023 and that the execution of the December Plan, including cash payments, will be substantially complete by the end of the first quarter of 2024. Potential position eliminations in each country are subject to local law and consultation requirements, which may extend this process beyond the first quarter of 2024 in certain countries. The charges that the Company expects to incur are subject to a number of assumptions, including local law requirements in various jurisdictions, and actual expenses may differ materially from the estimates disclosed above. As part of the December Plan, Twilio Flex, the Company’s cloud contact center, will be reported as part of the Company’s Twilio Communications reportable segment in future periods. Prior periods presented for purposes of comparison will be recast accordingly.

D. UniEval Implementation Details

Given the constraints of the context window in the model originally associated with the UniEval (Zhong et al., 2022) metric, we implemented an approach on GPT-4-turbo that focuses on the originally proposed dimensions: *coherence*, *consistency*, *fluency*, *relevance*, and *groundedness*. The questions were adapted in the following manner, in which *analyses* pertain to the text generated and *document* refers to the associated MF form:

Coherence Is this an analysis with ideas that are coherent with each other?

Consistency Is this an analysis consistent with the facts in the document?

Fluency Is this a fluent analysis?

Relevance Is this an analysis that covers the most relevant topics of the document?

Groundedness Does this analysis develop insights derived from the knowledge in the document?

Therefore, GPT-4 was required to respond to the mentioned questions with “Yes” or “No”. Subsequently, we used the *top_logprobs* API parameter to compute the formulations as originally proposed by the metric. Additionally, it is pertinent to underline that all the dimensions in use have their results ranging from 0 to 1.

E. Examining AI-generated vs. human-written analyses

This section highlights the similarities and differences observed between an AI-generated analysis and the human-written reference. The analyses are presented in Appendices E.1 and E.2, respectively; the specific MF form addressed by these texts is outlined in Appendix C.3, pertaining to the Petrobras case.

Foremost, the analysis by Mistral 7B is notably shorter, as previously indicated in the Table 2. Both texts rely on the events described in the MF. Specifically, the AI-generated text concentrates on detailing the facts presented in the form. Furthermore, the texts diverge in their coverage of certain topics. The AI-generated text focuses on the acquisition regime present in the MF and its potential implications (“*The Production Sharing Regime allows for the sharing of risks and rewards between the government and the private sector, which can encourage investment in exploration and production*”). On the other hand, the human-written text explores additional themes beyond just oil production potential increase, such as job creation, economic stimulus, and technological advancement.

Nonetheless, both texts address topics which are not explicitly mentioned in the MF. Notably, both point out that the company is under Brazilian state control, a potential important detail for contextualizing the information for readers. Similarly, each text acknowledges the company’s long-standing expertise in the region, with Mistral 7B noting, “*The basin has a long history of oil and gas production,*

and Petrobras has been a key player in the region for many years...” while the human analysis adds, “Petrobras, as the operator of the surrounding fields in the Campos Basin, has the necessary expertise and infrastructure for the swift development of the Jaspe block”. Both analyses discuss risks, though in different aspects. The AI-generated text highlights competition from other oil and gas companies and potential regulatory challenges. In contrast, the human-written text mentions actual issues like environmental concerns and fluctuating oil prices.

Nevertheless, while both cover important common points, the approach to certain topics varies. For instance, the AI-generated analysis superficially mentions, “The Jaspe block is located in the Campos Basin, which is among the most prolific oil and gas basins in Brazil”. It may be inferred from the MF that Petrobras’s strategy includes “...focusing on profitable assets and replenishing oil and gas reserves...”. Conversely, the human-written text offers a more quantitative perspective and is supported by additional data, with the statement, “...Campos Basin, a geological treasure trove responsible for about 80% of Brazil’s oil production. This offshore block is estimated to contain up to 3 billion barrels of oil equivalent...”.

In addition, another significant aspect is found in the recommendations sections of both texts. While the Mistral 7B concludes its analysis by revisiting the mentioned risks and pointing out the event as potentially positive move, the human-written reference brings in additional external context by evaluating other company alternatives for potential investment, ultimately recommending a “hold” stance.

Although here we refrain from determining which of the two texts addresses the event more effectively due to our lack of specialized knowledge in the finance sector, it is noteworthy that both texts present topics relevant to the MF, albeit from different styles. We acknowledge that these differences may be pivotal and significant in shaping a reader’s direct preference, and therefore justifies further investigation.

E.1. Analysis generated by Mistral 7B

Petrobras’ Interest in Jaspe Block Under Production Sharing Regime: An Analysis

Petrobras, the state-owned oil and gas company in Brazil, has recently expressed interest in the right of first refusal in the Jaspe block, located in the Campos Basin, under the Production Sharing Regime. This analysis will examine the material facts surrounding Petrobras’ interest in the Jaspe block and provide recommendations for investors. Petrobras has expressed interest in the right of first refusal in the Jaspe block, located in the Campos Basin, under the Production Sharing Regime. The

expression of interest is in line with Petrobras’ E&P strategy set out in SP 24-28+, focusing on profitable assets and replenishing oil and gas reserves.

Petrobras’ interest in the Jaspe block is significant as it represents a potential opportunity for the company to increase its oil and gas reserves and maintain its position as a major player in the Brazilian energy sector. The Production Sharing Regime allows for the sharing of risks and rewards between the government and the private sector, which can encourage investment in exploration and production.

The Jaspe block is located in the Campos Basin, which is one of the most prolific oil and gas basins in Brazil. The basin has a long history of oil and gas production, and Petrobras has been a major player in the region for many years. The block is estimated to have significant oil and gas reserves, which could provide a significant boost to Petrobras’ E&P activities.

Recommendation: Given Petrobras’ interest in the Jaspe block and the potential benefits of the Production Sharing Regime, investors should closely monitor the company’s activities in the region. Petrobras’ E&P strategy is focused on profitable assets and replenishing oil and gas reserves, which could lead to increased production and revenue in the long term.

Investors should also be aware of any potential risks associated with Petrobras’ interest in the Jaspe block, such as competition from other oil and gas companies in the region and potential regulatory challenges. However, given Petrobras’ long history of success in the Brazilian energy sector and its focus on profitable assets, investors should view the company’s interest in the Jaspe block as a positive development.

Conclusion: Petrobras’ interest in the Jaspe block under the Production Sharing Regime represents a potential opportunity for the company to increase its oil and gas reserves and maintain its position as a major player in the Brazilian energy sector. Investors should closely monitor the company’s activities in the region and be aware of any potential risks associated with Petrobras’ interest in the Jaspe block. Overall, Petrobras’ focus on profitable assets and replenishing oil and gas reserves could lead to increased production and revenue in the long term.

E.2. Human-written Analysis

Petrobras (PBR) Plans to Acquire Jaspe Block in Offshore Brazil¹⁰

Petrobras PBR, the leading Brazilian state-run oil company, announced its intention to acquire the

¹⁰<https://bit.ly/zacks-pbr-jaspeblock>

Jaspe oil block, located in the Campos Basin, by exercising its preferential rights. The offshore block is expected to be tendered by the country's regulator. The Jaspe block holds considerable significance for Brazil's oil production, and Petrobras anticipates substantial gains from the same. The block is estimated to contain significant oil reserves, which will help Petrobras increase its production capacity. Let's delve deeper into this strategic acquisition and explore its implications.

Jaspe Block: A Jewel in the Campos Crown

The Jaspe block sits nestled within the heart of the Campos Basin, a geological treasure trove responsible for roughly 80% of Brazil's oil output. This offshore block holds estimated reserves of up to 3 billion barrels of oil equivalent, making it a highly attractive prospect for any oil company. Its shallow waters and proximity to existing infrastructure further enhance its appeal, promising lower development costs and faster production timelines.

Petrobras' Strategic Plan: Unlocking Potential

Petrobras, as the operator of the surrounding fields in the Campos Basin, boasts the necessary expertise and infrastructure to swiftly develop the Jaspe block. This acquisition aligns perfectly with the company's strategic focus on maximizing output from mature basins while pursuing exploration in frontier areas. By leveraging its existing assets and knowledge, Petrobras can unlock the potential of the Jaspe block.

Increased Oil Production: Jaspe's reserves hold the potential to significantly boost Petrobras' production, bolstering its position as a key player in the global oil market. This translates to enhanced energy security and economic benefits for Brazil.

Job Creation and Economic Stimulus: Developing the Jaspe block will inevitably create new job opportunities, both directly in the oil and gas sector and indirectly through supporting industries. This economic stimulus will ripple through local communities, fostering growth and development.

Technological Advancement: Extracting oil from the Jaspe block will necessitate cutting-edge technologies and innovative solutions. This fosters a robust research and development environment, propelling Brazil's oil industry forward.

A Win for Brazil's Oil Ambitions. The Jaspe acquisition isn't just a win for Petrobras, it serves as a vital stepping stone for Brazil's broader oil production goals. The successful development of this block sends a positive signal to international investors, potentially attracting further investment and expertise to the Brazilian oil sector. This fosters healthy competition and accelerates exploration and development activities, propelling Brazil toward ambitious oil production targets.

Conclusion: Challenges and Considerations: While the Jaspe block holds immense potential, it's

not without challenges. Environmental concerns surrounding offshore drilling activities require careful mitigation strategies. Additionally, fluctuating oil prices and global economic shifts can impact the project's viability. Petrobras must navigate these challenges prudently to ensure long-term success of the Jaspe acquisition.

A Strategic Move With Far-Reaching Implications Petrobras' plan to acquire the Jaspe block marks a strategic move with the potential to significantly benefit both the company and Brazil's oil industry as a whole. Increased production, economic stimulus and technological advancements are just some of the rewards on the horizon. While challenges remain, the successful development of Jaspe could unlock a new chapter in Brazil's oil production story, solidifying its position as a major player in the global energy landscape.

Recommendation: Currently, PBR carries a Zacks Rank #3 (Hold). Investors interested in the energy sector might look at some better-ranked stocks like Sunoco LP SUN and Oceaneering International, Inc. OII, both sporting a Zacks Rank #1 (Strong Buy), and Enbridge Inc. ENB, carrying a Zacks Rank #2 (Buy) at present. You can see the complete list of today's Zacks #1 Rank stocks [here](#).