Risks of AI-Assisted Forecasted Inputs in Discounted Cash Flow Valuation: Reproducibility and Transparency

Romana Čižinská[1] – Pavel Neset[2]
ORCID iD: 0000-0003-2313-5760[1], 0000-0002-0643-0162[2]
romana.cizinska@savs.cz, pavel.neset@savs.cz
[1] Skoda Auto University, Department of Finance and Accounting,
Mlada Boleslav, Czech Republic
[2] Skoda Auto University, Department of Law and Economics,
Mlada Boleslav, Czech Republic

https://doi.org/10.18267/pr.2026.vol.2587.7

Abstract: The growing use of large language models (LLMs) to generate forecasted inputs in discounted cash flow (DCF) valuation raises important methodological concerns related to reproducibility and transparency. DCF valuation critically depends on forecasts of key input variables. According to the principles of the International Valuation Standards (IVS), valuation conclusions must be based on a documented process that allows independent reproduction of conclusions under identical assumptions. The aim of this paper is therefore to examine whether AI-assisted forecasts of valuation-relevant inputs remain consistent when identical tasks are repeated, and to assess the implications of observed variability for the reproducibility of the valuation process. Based on an experiment involving repeated runs of an identical prompt, the study shows that differences arise not only in the wording of the responses but also in point estimates, forecast intervals, and the accompanying reasoning. The findings suggest that AI-assisted forecasting of valuation inputs cannot be considered fully reproducible in the sense required by IVS without additional control mechanisms. The paper therefore proposes the archiving of prompts, outputs, and metadata, repeated runs of identical tasks, and the explicit separation of factual inputs from expert judgement as minimum conditions for transparent and reviewable use of AI-assisted forecasts in DCF valuation.

Keywords: international valuation standards (IVS), reproducibility, transparency, AI-assisted forecasted inputs, discounted cash flow (DCF) valuation

JEL Classification codes: G17, C52

INTRODUCTION

With the rapid development of generative artificial intelligence based on large language models (LLMs), these tools are beginning to find applications in income-based business valuation. Their use is not limited to supporting textual reasoning in expert or analytical reports; increasingly, they also influence the preparation and interpretation of the key value drivers used in income-based valuations. The income approach, applied through the discounted cash flow (DCF) method is based on analysing and forecasting the fundamental value drivers, i.e., revenues and revenue growth, profit margins, investments in fixed assets and net working capital, and the discount rate. Because the future is not objectively fully knowable and any estimate is inherently subject to uncertainty, professional judgement, and the limitations of available information, forecasting these variables represents one of the most methodologically demanding parts of the valuation.

In this respect, the advantages of LLMs lie particularly in their speed, their ability to organise large volumes of information, and their ability to generate plausible-sounding explanations of the assumptions used. In some cases, such AI-generated estimates may even be numerically close to the outcome that is later realised. At the same time, however, the growing use of LLMs raises important methodological questions for income-based valuation. The fact that an output appears expert, consistent, or ex post approximates the realised outcome does not in itself mean that such an output is suitable for a specific valuation engagement whose objective is to determine a particular basis of value for a specific subject, asset, and valuation purpose as of a given valuation date. In professional valuation, it is not sufficient to merely "arrive at a number." Equally important is the ability to transparently demonstrate the information set used, the assumptions applied, and the process by which the valuer arrived at the conclusion of value. Only such a process supports the reviewability of the valuation conclusion and provides a basis for another independent valuer to reach a comparable conclusion when applying the same information set, assumptions, and methodology. In this sense, the notion of valuation reproducibility can be understood.

The requirement for transparency and reviewability of the valuation process is also implicitly embedded in the International Valuation Standards (IVS). IVS 100, paragraphs 20.01 and 20.02 (International Valuation Standards Council [IVSC], 2024) state that the valuation process must be conducted objectively, transparently, and without bias, and that appropriate valuation process quality controls must be in place. These controls must also be documented in sufficient detail to enable another valuer, applying professional judgement, to understand how they operate and to assess the appropriateness of the procedures and judgements used (IVS 100, paragraph 20.05 in IVSC, 2024). Transparency of the valuation process is further supported by requirements concerning the scope of work and sources of information. The scope of work for a valuation that is appropriate for the intended use must explicitly specify the nature and sources of significant information relied upon by the valuer, as well as the extent of analyses performed and the nature of documentation provided (IVS 101, paragraph 20.01 in IVSC, 2024). For each valuation approach applied, the standards also require the valuer to analyse the inputs and assumptions used in the valuation and assess their appropriateness with respect to the valuation purpose (IVS 103, paragraph 10 in IVSC, 2024). All significant assumptions must be reasonable under the circumstances and supported by available information (IVS 102, paragraph 50.04 in IVSC, 2024). Moreover, the valuer should be aware that most bases of value prohibit the consideration of information or market sentiment that would not be known or knowable to participants with reasonable due diligence on the valuation date (IVS 102, paragraph 10.04 in IVSC, 2024).

IVS do not provide detailed methodological rules for generating forecasts of the input variables used in income-based valuation, nor for assessing the tools that may be employed to produce those forecasts. This opens the question of how these principles should be applied when forecasts are generated using generative AI tools.

The aim of this paper is to determine whether forecasts of input variables for income-based valuation generated by large language models differ when the same task is repeated, and subsequently to assess what implications any such variability has for the reproducibility of income-based valuation. Naturally, the requirement of reproducibility is not the only condition for an IVS-compliant valuation, but it constitutes an important component of the broader requirement for a transparent, reviewable, and professionally defensible valuation process consistent with IVS principles. Other relevant conditions for the applicability of AI-assisted income-based valuation are the subject of our broader research and are developed in related papers.

1 LITERATURE REVIEW

Large language models (LLMs) are not fully deterministic in their outputs. Even when given identical input, they may not always produce the same response, reflecting both the probabilistic nature of the generative process and the inherent stochasticity of LLM outputs. This phenomenon has also been confirmed empirically.

Ouyang et al. (2025) analyse the non-determinism of ChatGPT in code generation. Using a large set of programming tasks, they show that the model often produces different outputs, resulting in low reproducibility. Ouyang et al. (ibid.) note that researchers need to take non-determinism into account when drawing conclusions in order to place LLM-based research on firmer scientific foundations. Similarly, Atil et al. (2025) examine LLM non-determinism in settings that are typically assumed to be deterministic and show that repeated runs may differ in textual form of the answer, even though the underlying extracted answer remains the same. Pinhanez et al. (2025) likewise find that repeated attempts to answer identical multiple-choice questions may produce different answers, with relatively low run-to-run agreement. Nicholson (2026) reaches similar conclusions and argues that such differences may also stem from technical aspects of model operation, including numerical computation, parallelisation of inference, and the distribution of model instances across processes. These factors contribute to what the literature refers to as "baseline drift" or "output drift", a phenomenon also documented in tasks with a formally defined output structure (e.g., Khatchadourian, 2026).

Beyond the general literature on LLM output variability, studies specifically examining applications in finance and accounting have begun to emerge. Wang and Wang (2025) analyse the consistency and reproducibility of LLM outputs when working with financial texts and conduct an extensive experiment comprising 50 independent runs of identical prompts. Their findings show that reproducibility depends strongly on the task type. While simple classification tasks exhibit very high consistency across runs, more complex tasks — such as text generation or numerical prediction — show substantially higher variability. Yu (2023) similarly highlights the importance of output variability, analysing the stability of sentiment classification in financial news. The study demonstrates that repeated queries can lead to significant differences in assigned sentiment. These discrepancies propagate into investment strategies that rely on sentiment signals — producing different portfolios and different investment returns even when using the same model and the same dataset. Lopez-Lira and Tang (2023) further show that LLM-generated sentiment can be used to predict stock price movements, suggesting that LLM outputs may have direct economic consequences for investment decisions. Recent studies also show that variability does not occur only in creative tasks but also in tasks where identical outputs would, in principle, be expected. Cheung (2024) demonstrates that when answering the same data-analysis question, an LLM may produce different solution paths across runs, leading to variation in output quality (e.g., accuracy, match rate, run rate). In contrast, Kim et al. (2024) examine the use of LLMs in financial analysis but do not explicitly test for run-to-run variability. Overall, the available evidence suggests that variability may occur even in tasks with a fixed structure and an expected unique output.

The literature further indicates that output variability is particularly important in regulated financial services. Khatchadourian (2026) notes that LLM variability may complicate audit replay in regulated settings, where a regulator may require institutions to reproduce a specific system decision when a model is re-executed with identical inputs — an expectation that presupposes a high degree of output determinism.

In the context of this paper, existing research provides relatively robust evidence that LLM outputs can exhibit variability across repeated identical prompts. This variability is not limited to generative or creative tasks but may also affect predictive and quantitative tasks, including economic estimation. Some studies further show that variability can influence investment strategy performance (Yu, 2023), underscoring its practical relevance. For income-based

valuation, this implies that model outputs may differ across runs to an extent that is both statistically and practically significant.

The observed variability is attributed to several complementary mechanisms. One is the sampling-based nature of LLM generation: models do not produce outputs as fully deterministic systems but select successive tokens based on probability distributions, meaning that even identical prompts may yield slightly different answers. Empirical research also shows that variability may occur even when the temperature parameter is set to zero — a configuration commonly assumed to be deterministic (Ouyang et al., 2025; Atil et al., 2025). Beyond the probabilistic nature of generation, the literature highlights the role of technical infrastructure, such as floating-point arithmetic differences, parallel processing, and implementation details of the inference environment (Nicholson, 2026). Models are also sensitive to small variations in prompt wording and context (prompt sensitivity), which can meaningfully affect outputs (Herrera Poyatos et al., 2025). Recent synthetic surveys of financial LLMs further identify output stability, reliability, and consistency as major challenges for financial applications (Lee et al., 2025; Nie et al., 2024). In addition to output variability, the literature highlights another methodological limitation relevant for economic and financial forecasting: the risk of memorisation of historical data and associated look-ahead bias, whereby the model may use information that would not have been available at the valuation date (Lopez-Lira et al., 2025).

Although the literature explains why LLM outputs may vary — whether due to stochastic generation, prompt sensitivity, technical properties of model execution, or look-ahead bias — it does not yet provide adequate guidance on how this property should be methodologically addressed when AI-generated forecasts are used as input assumptions in income-based business valuation that must withstand IVS requirements for transparency, reviewability, and reproducibility of the valuation process. There is currently no systematic methodological literature addressing how such variability should be incorporated into forecasting the input variables used in DCF valuation, or how to handle it in light of IVS. To date, the issue has been acknowledged only marginally, and the literature does not offer more comprehensive methodological frameworks linking AI-generated outputs to IVS requirements for reproducible valuation processes. This gap is particularly significant because reproducibility is one of the key prerequisites for a professionally defensible and standards-compliant income-based valuation.

2 METHODOLOGY

This paper employs an experimental research design to analyse the reproducibility of AI assisted forecasting of valuation inputs in the context of income based valuation using the DCF method. The aim of the experiment is to determine whether identical prompts executed under identical input conditions produces identical (or at least highly similar) outputs, or whether the model generates variable results that may have implications for the reproducibility of AI assisted inputs within the valuation process.

DCF valuation context and the role of the macroeconomic input

The experiment is conducted within the valuation context of an unlisted engineering company, ABC, a.s., operating in the Czech Republic and specialising in the production of components and technological units for the automotive and industrial sectors. The subject of the valuation is a 100% equity share in the company for the purposes of litigation. The relevant basis of value is Market Value, i.e. the estimated amount for which an asset and/or liability should exchange on the valuation date (31 December 2023) between a willing buyer and a willing seller in an arm’s-length transaction, after proper marketing and where the parties had each

acted knowledgeably, prudently and without compulsion (see IVS 102, paragraph A10.01 in IVSC, 2024).

Given the nature of the business (an unlisted company) and the limited applicability of comparable transactions or publicly traded peers, the income approach — specifically the DCF entity method — was selected as the primary valuation approach. Enterprise value is derived from the nominal expected free cash flows to the firm (FCFF) and the corresponding nominal discount rate (WACC) determined from the perspective of a hypothetical market participant as of the valuation date. Equity value is then obtained as enterprise value adjusted for interest-bearing debt and cash as of the valuation date. Based on the site visit, it was established that the revenues of ABC, a.s. are influenced both by a volume component linked to customer production levels and a price component, which is indexed to inflation in part of the contracts. The company's cost base is materially affected by labour costs and other operating expenses, where inflation also feeds into their development. The forecasted macroeconomic variable is the Harmonised Index of Consumer Prices (HICP), expressed as a year-over-year rate for the 12-month period following the valuation date. Within a nominal DCF model, HICP serves as an important macroeconomic input that anchors the inflation component of the projections. The forecasted HICP does not enter the valuation as a standalone variable but indirectly through its role in constructing nominal free cash flows to the firm (FCFF).

Experimental design and limitations

The objective of the experiment is to analyse the reproducibility of AI-assisted macroeconomic forecasts under identical input conditions. The experiment consists of repeatedly running an identical prompt in the environment of the generative language model ChatGPT. To illustrate the variability of AI-generated outputs, the same prompt is executed 15 times in separate runs (new chats), with all runs conducted within the same time window. This number of repetitions was chosen to provide a sufficient basis for observing potential variability in the outputs while remaining feasible within the experimental design. Repeated runs allow for a straightforward quantitative assessment of the dispersion in the resulting point estimates of HICP inflation.

The prompt is formulated to comply with the methodological requirements of valuation under IVS, particularly the requirement that the information used must be "known or knowable" as of the valuation date. The model is instructed to estimate inflation exclusively on the basis of a predefined set of documents (data-pack) available as of the valuation date (31 December 2023) and to refrain from using any other sources of information. The prompt is formulated to reflect IVS requirements regarding time-specificity and control over the information set at the relevant date. It includes a definition of the valuation context (DCF valuation), an explicit statement of the valuation date, a restriction of the information set to documents published before that date, a prohibition on using other sources or the model's implicit knowledge, a requirement to cite sources via their DocID identifiers, and an obligation to separate factual statements from professional judgements. The instructions also require the creation of an audit trail at the end of the response. The model is instructed to provide an "AI output record" at the end of each answer, containing the model identifier, the model version (if available), the time of generation (if accessible to the model), and a declaration that no sources outside the defined data-pack were used in constructing the response. If any of these items are not available to the model, it is instructed to explicitly state "not available to the model." This requirement is intended to strengthen the audit trail and document the context in which the output was generated.

The data-pack used in the experiment contains three macroeconomic sources published before 31 December 2023: (1) European Commission: European Economic Forecast, Autumn 2023 (Doc01), (2) IMF: World Economic Outlook, October 2023 (Doc02), and (3) OECD: Economic Outlook, Volume 2023 Issue 2 (Doc03). These documents represent typical publicly available

sources of macroeconomic forecasts that a hypothetical market participant might use when forming expectations as of the valuation date.

The aim of the experiment is not to test the predictive ability of the model as such, but to analyse whether identical input leads to identical or divergent outputs, and thus whether AI-assisted forecasts can be considered fully reproducible. The requirement to include the AI output record serves to reinforce the transparency and reviewability of the output-generation process. From a methodological perspective, this element provides a practical operationalisation of IVS requirements regarding documentation of the inputs used and enables subsequent identification of the context in which a given AI-assisted estimate was produced — a factor that is essential for assessing the reproducibility and auditability of the valuation process.

Although the experiment provides insights into the variability of AI-generated forecasts, several limitations should be acknowledged. First, the analysis is based on a limited number of repeated separate runs of the identical prompt. While sufficient to illustrate output variability, a larger number of runs could provide a more robust statistical assessment. Second, the experiment focuses on a single large language model (ChatGPT), and the observed variability may differ across models or model configurations. Third, the analysis examines a single valuation-relevant macroeconomic input (HICP inflation). The extent of variability may differ for other inputs commonly used in DCF valuation, such as revenue growth, operating margins, or discount rate components.

Evaluation of Outputs Across Individual Runs

The outputs of the fifteen repeated runs are evaluated in terms of their numerical variability, in particular by comparing the point estimates and forecast intervals of HICP across individual executions of the model. Attention is given to whether, under identical input conditions, only the linguistic formulation of the response changes, or whether differences also appear in the numerical inflation estimate itself and in the reasoning provided. In addition to numerical discrepancies, the analysis therefore also examines the consistency of the main factors on which the model bases its projections. To provide a basic quantitative description of variability, simple descriptive statistics — specifically the mean, median, and sample standard deviation — are computed for the point estimates.

To illustrate the economic relevance of the observed variability, the AI-generated estimates are also compared with the ex post realised average annual HICP inflation rate for 2024, corresponding to the twelve month horizon following the valuation date. This comparison, however, serves only as a supplementary reality check and not as the primary criterion for evaluating the individual runs. The main focus of the analysis is the reproducibility of AI-assisted outputs under identical input conditions (namely, whether an identical prompt yields identical or at least highly similar results).

The evaluation also includes an assessment of the procedural transparency of the outputs, particularly whether the model complies with the requirement to separate facts from judgements, whether it provides identifiable sources (DocID), and whether the logic of the estimate can be reconstructed from the stated arguments.

The archiving of the prompt, individual responses, and accompanying metadata creates an audit trail that enables systematic comparison across runs and supports an assessment of whether AI-assisted inputs to the valuation process can be considered sufficiently reproducible and reviewable.

3 RESULTS AND DISCUSSION

Table 1 summarises the outputs of the AI-assisted forecast of HICP (year-on-year) for the next 12 months as of the valuation date 31 December 2023, generated in fifteen independent repetitions of an identical prompt executed in new, separate chat sessions. All runs were conducted in the ChatGPT environment (version 5.3) under a single user account. In the first round, three runs were classified as unsuccessful because they did not provide a fully usable numerical output in the specified format. In one case, the model concluded the response with the statement "insufficient evidence" without providing a clear final quantitative estimate, and in two additional cases the obtained outputs could not be unambiguously classified within the comparable structure of point and interval estimates. For this reason, these runs were excluded from the final evaluation and replaced with new independent repetitions so that the final dataset contained 15 fully comparable outputs.

Tab. 1 Variability of HICP inflation forecasts across repeated separate runs of the identical prompt (as of 31 December 2023)

Run	Point estimate (%)	Lower bound (%)	Upper bound (%)	Main stated drivers
1	3.5	2.5	4.5	Fading energy shocks, restrictive monetary policy, gradual disinflation
2	3.0	2.0	4.0	Post-energy-shock disinflation, monetary tightening, normalization of energy prices
3	3.0	2.5	3.5	Monetary tightening, declining energy prices, weaker demand
4	2.8	2.3	3.3	Restrictive monetary policy, fading energy shocks, wage pressures
5	3.5	2.5	4.5	Monetary tightening, normalization of supply chains, fading price shocks
6	2.6	2.0	3.2	Declining energy prices, restrictive monetary policy, persistent services inflation
7	3.0	2.5	3.5	Monetary tightening, normalization of supply chains, declining commodity prices
8	3.5	2.8	4.2	Fading energy shocks, restrictive monetary policy, persistent core inflation
9	3.0	2.0	4.5	Monetary tightening, energy price volatility, wage pressures
10	2.9	2.4	3.3	Energy price stabilization, monetary tightening, services inflation
11	3.0	2.5	3.5	Monetary tightening, normalization of commodity prices, gradual disinflation
12	3.2	2.5	4.0	Fading price shocks, monetary tightening, base effects
13	2.8	2.0	3.5	Declining energy prices, restrictive monetary policy, services inflation
14	4.0	3.0	5.0	Slower disinflation, wage pressures, persistent services inflation

Run	Point estimate (%)	Lower bound (%)	Upper bound (%)	Main stated drivers
15	2.7	2.0	3.5	Energy price normalization, monetary tightening, gradual disinflation

Source: Authors' own experiment based on repeated AI-assisted forecasts (ChatGPT), using macroeconomic sources available as of 31 December 2023 (European Commission, IMF, OECD)

The results presented in Table 1 show that even when an identical prompt is used, the resulting forecasts are not identical, despite the fact that the valuation task and the type of valuation relevant input remain the same. The repeated runs produced outputs that differed not only stylistically but also substantively. As shown in Table 1, variability manifested itself in the point estimate, in the width of the forecast interval, and in partial differences in the justification of the projection. The distribution of point estimates across the 15 independent runs shows a mean forecast of approximately 3.1%, a median of 3.0%, and a sample standard deviation of 0.38 percentage points, indicating meaningful variability in AI generated forecasts even under identical prompting conditions. The estimated lower bounds range from 2.0% to 3.0%, with a mean of approximately 2.4%, while the upper bounds range from 3.2% to 5.0%, with a mean of roughly 3.8%. The dispersion of both lower and upper bounds further indicates that repeated prompting affects not only the central estimate itself but also the model's implied uncertainty about future inflation. These differences therefore do not merely reflect alternative verbal formulations of the same conclusion; they represent numerically distinct and economically relevant outcomes accompanied by not fully identical reasoning.

For reference, according to official Eurostat data, the Czech Republic recorded an average annual HICP inflation rate of 2.7% in 2024 (Eurostat, 2026). This realised value falls within the forecast range generated in most of the repeated runs and lies close to the lower part of the distribution of point estimates reported in Table 1. This comparison serves solely as a supplementary reality check and does not alter the main conclusion regarding the variability of AI generated forecasts.

The observed variability is consistent with the existing literature on the properties of large language models. Empirical studies repeatedly show that LLMs are not fully deterministic and that repeated queries using an identical prompt do not necessarily yield identical outputs. The fact that fifteen identical queries did not result in fifteen identical forecasts is therefore not anomalous; it aligns with a well documented characteristic of these models (Ouyang et al., 2025; Atil et al., 2025).

At the same time, the nature of the task must be taken into account. Forecasting future inflation is a predictive task involving economic interpretation and numerical estimation. Empirical studies show that precisely these types of tasks tend to exhibit higher output variability than simpler classification or extraction tasks (Wang & Wang, 2025). One cannot rule out the influence of technical factors relating to the model's execution either. The literature notes that even with identical input, minor numerical differences in computation or implementation characteristics of the inference infrastructure may lead to slightly different sequences of generated tokens and thus different responses (Nicholson, 2026). These mechanisms — sometimes referred to as output drift or baseline drift — represent additional possible explanations for the observed discrepancies across runs.

The experiment demonstrates not only that LLM outputs exhibit variability, but also that this variability persists even under a tightly controlled valuation task that respects the valuation date, a restricted information set, and explicit auditing instructions. The experiment thus shows that even relatively strong prompt standardisation does not, by itself, ensure full reproducibility of an AI assisted forecast. The prompt alone therefore cannot be considered a sufficient carrier

of valuation methodology in the sense of the IVS requirements for transparency, reviewability, and reproducibility of the valuation process. From this perspective, an AI generated forecast should not be understood as an autonomously usable valuation input, but rather as an analytical support whose role, scope, and mode of use must be predefined in the scope of work and complemented by appropriate control mechanisms. These include, in particular, archiving the exact prompt, documenting the specific output used in the valuation, performing repeated runs of the identical input to assess output variability, documenting the valuer's selection of the final estimate, and explicitly separating factual evidence from professional judgement. Reproducibility in AI assisted valuation therefore cannot be based on the expectation that the model will always return the same output for the same input, but rather on a transparently documented process that allows variability to be identified, evaluated, and methodologically addressed.

CONCLUSION

This paper demonstrates that AI-assisted generation of input assumptions for income-based (DCF) valuation brings, alongside potential benefits, significant methodological risks related to the reproducibility of results. The experiment involving repeated runs of an identical prompt showed that, even under unchanged input conditions, a generative model does not necessarily produce identical outputs. The variability manifested not only in the stylistic form of the responses but also in the numerical point estimates, the width of forecast intervals, and, to some extent, in the reasoning underlying the outputs. In the context of DCF valuation, even a relatively small deviation in a macroeconomic input can cascade through other dependent assumptions in the model, ultimately affecting the defensibility of the final conclusion of value.

From a valuation-methodology perspective, this finding is particularly important in relation to the requirement of reproducibility of the valuation process. If identical input leads to differing outputs, AI-assisted inputs cannot, without additional measures, be considered fully reproducible. Under IVS, what matters is not only the forecasted figure itself but also the ability to retrospectively reconstruct the process, information set, and professional judgements on which that figure was based. Output variability in AI-generated estimates therefore represents not merely a technical characteristic of the model but a methodologically relevant issue for transparency, reviewability, and auditability of income-based valuation.

The findings further indicate that reproducibility in AI-assisted valuation cannot be ensured simply by standardising the prompt. It is necessary to complement the use of LLMs with procedural mechanisms that allow this variability to be transparently documented and professionally evaluated. The proposed approach — based on the archiving of prompts and outputs, repeated runs of identical queries, and the explicit separation of factual inputs from professional judgement — provides a practical operationalisation of IVS requirements concerning transparency, reviewability, and defensibility of the valuation process. In this way, the paper contributes to the literature by illustrating how the well-documented tendency of LLMs to produce variable outputs under identical input conditions translates into the specific context of income based business valuation and its methodological requirements under IVS. This study is subject to several limitations, including the focus on a single LLM, a limited number of repeated runs, and the analysis of only one valuation relevant input variable. Other inputs used in DCF valuation may exhibit different patterns of variability. Future research could extend the analysis to additional inputs, models, and prompt designs, with particular emphasis on developing procedures that ensure transparency, reviewability, and reproducibility of LLM assisted valuation in line with the principles of IVS.

REFERENCES

Atıl, B., Aykent, S., Chittams, A., Fu, L., Passonneau, R. J., Radcliffe, E., Rajagopal, G. R., Sloan, A., Tudrej, T., Ture, F., Wu, Z., Xu, L., & Baldwin, B. (2025). Non-determinism of "deterministic" LLM system settings in hosted environments. In *Proceedings of the 5th Workshop on Evaluation and Comparison of NLP Systems* (pp. 135–148). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.eval4nlp-1.12

Cheung, M. (2024). A reality check of the benefits of LLM in business [Preprint]. *arXiv*. https://arxiv.org/abs/2406.10249

Eurostat. (2026). *Harmonised index of consumer prices (HICP): ECOICOP ver. 2 - indices and rates of change, annual data*. Retrieved April 15, 2026, from https://ec.europa.eu/eurostat/databrowser/view/prc_hicp_ainr/default/table?lang=en

Herrera-Poyatos, D., Peláez-González, C., Zuheros, C., Herrera-Poyatos, A., Tejedor, V., Herrera, F., & Montes, R. (2025). An overview of model uncertainty and variability in LLM-based sentiment analysis: Challenges, mitigation strategies, and the role of explainability. *Frontiers in Artificial Intelligence, 8*, Article 1609097. https://doi.org/10.3389/frai.2025.1609097

International Valuation Standards Council. (2024). *International Valuation Standards (IVS)*. https://ivsc.org/standards/

Khatchadourian, R. (2026). Replayable financial agents: A determinism-faithfulness assurance harness for tool-using LLM agents [Preprint]. *arXiv*. https://arxiv.org/abs/2601.15322

Kim, A. G., Muhn, M., & Nikolaev, V. V. (2024). Financial statement analysis with large language models [Preprint]. *arXiv*. https://arxiv.org/abs/2407.17866

Lee, J., Stevens, N., & Han, S. C. (2025). Large language models in finance (FinLLMs): A survey. *Neural Computing and Applications, 37*(30), 24853-24867. https://doi.org/10.1007/s00521-024-10495-6

Lopez-Lira, A., & Tang, Y. (2023). Can ChatGPT forecast stock price movements? Return predictability and large language models [Preprint]. *arXiv*. https://arxiv.org/abs/2304.07619

Lopez-Lira, A., Tang, Y., & Zhu, M. (2025). The memorization problem: Can we trust LLMs' economic forecasts? [Preprint]. *arXiv*. https://arxiv.org/abs/2504.14765

Nicholson, C. (2026). Quantifying non-deterministic drift in large language models [Preprint]. *arXiv*. https://arxiv.org/abs/2601.19934

Nie, Y., Kong, Y., Dong, X., Mulvey, J. M., Poor, H. V., Wen, Q., & Zohren, S. (2024). A survey of large language models for financial applications: Progress, prospects and challenges [Preprint]. *arXiv*. https://arxiv.org/abs/2406.11903

Ouyang, S., Zhang, J. M., Harman, M., & Wang, M. (2025). An empirical study of the non-determinism of ChatGPT in code generation. *ACM Transactions on Software Engineering and Methodology, 34*(2), Article 42. https://doi.org/10.1145/3697010

Pinhanez, C., Cavalin, P., Santos, C., Grave, M., & Primerano, Y. (2025). The non-determinism of small LLMs: Evidence of low answer consistency in repetition trials of standard multiple-choice benchmarks [Preprint]. *arXiv*. https://arxiv.org/abs/2509.09705

Wang, J. J., & Wang, V. X. (2025). Assessing consistency and reproducibility in the outputs of large language models: Evidence across diverse finance and accounting tasks [Preprint]. *arXiv*. https://arxiv.org/abs/2503.16974

Yu, B. (2023). Benchmarking large language model volatility [Preprint]. *arXiv*. https://arxiv.org/abs/2311.15180