

AI-Assisted Forecasts in Discounted Cash Flow Valuation: Valuation Date, Information Set, and Audit Trail

Romana CIZINSKA *

Skoda Auto University, Department of Finance and Accounting, Mlada Boleslav, Czech Republic

*Corresponding author, romana.cizinska@savs.cz

Pavel NESET

Skoda Auto University, Department of Law and Economics, Mlada Boleslav, Czech Republic

pavel.neset@savs.cz

Abstract. *This paper examines the growing use of generative artificial intelligence based on large language models in business valuation and related decision making. The scientific literature highlights efficiency gains in structuring information and developing scenarios, but also methodological risks when these tools are used to form valuation assumptions. A key concern is temporal leakage, also known as look-ahead bias, in which outputs may be influenced by information that was not available at the valuation date. Another concern is unsupported factual statements that cannot be traced to verifiable sources. These issues matter because, under International Valuation Standards, the conclusion on value must be stated as of a specific valuation date and must rely only on information that was known or knowable at that date. The paper investigates whether simply instructing a model to reason as of the valuation date is sufficient to prevent temporal leakage. It also examines how control over the information set affects the defensibility, reproducibility, and auditability of AI-assisted assumptions. It uses an illustrative discounted cash flow case and focuses on inflation measured by the harmonised index of consumer prices as an input to nominal projections of free cash flow to the firm. Three scenarios are compared: an open information window, an instruction-only regime without a closed information set, and a closed information window regime that restricts the model to a predefined data pack and requires a documented audit trail with source-linked factual claims and explicit reporting of insufficient evidence where support is missing.*

Keywords: Valuation Date, Temporal Leakage, Known Or Knowable Information Set, Audit Trail, Large Language Models, International Valuation Standards, Discounted Cash Flow Valuation.

Introduction

Generative artificial intelligence based on large language models (LLMs) is increasingly being used in income-based business valuation. Its application extends beyond drafting narrative reasoning and supporting documentation and is increasingly used in the analysis and forecasting of key value drivers such as revenues and revenue growth, profit margins, capital expenditure and net working capital, and the discount rate. Generative AI can undoubtedly structure information quickly, develop internally consistent scenarios, and describe the logic of assumptions in a transparent manner. At the same time, the associated methodological risks cannot be overlooked. Generated outputs may appear highly persuasive and sometimes even numerically accurate. However, this alone does not guarantee that they are suitable for real-world valuation engagements, where assumptions must be consistent with the specific basis of value, subject of valuation, and valuation (measurement) date. According to the International Valuation Standards Council (IVSC, 2024), the valuation date is the point in time to which the valuation applies.

For most valuation purposes, it is essential that the conclusion on value is stated as of the valuation date (IVS Glossary, paras. 10.30 and 10.32, in IVSC, 2024) and is based only on

information that was known or knowable at that date with reasonable due diligence (IVS 102, para. 10.04, in IVSC, 2024). This is particularly important in historical or retrospective valuations. As stated in IVS 102, paragraph 10.04 (IVSC, 2024), 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 with reasonable due diligence on the measurement/valuation date by market participants. In this context, the conceptual framework of market value further emphasises that prudence is assessed by referring to the state of the market at the valuation date, not with the benefit of hindsight at some later date (IVS 102, para. 10.02, in IVSC, 2024).

For the purposes of this paper, we introduce the concept of an information window, understood in two related senses. First, it denotes the temporal boundary that requires valuation assumptions to be derived only from information available as of the valuation date. Second, it denotes a specified information set (data pack), that is, a defined set of documents and data from which the assumptions are derived, together with rules and controls designed to prevent the inclusion of information that arose or became available only after the valuation date. Closing the information window is a natural precondition for the defensibility and reproducibility of the valuation outcome and should be reflected in the scope of work, which describes the fundamental terms of a valuation or valuation review (IVS 101, in IVSC, 2024).

When LLMs are used, these assumptions of defensibility and reproducibility can be breached in ways that may not be obvious at first glance. The risk is not limited to the inclusion of plausible-sounding but unsupported statements (hallucinations). The benefit of hindsight after the valuation date can also allow the use of information about subsequent developments, resulting in temporal leakage or look-ahead bias. Such contamination may lead to a hindsight-biased estimation of key DCF inputs and, consequently, to a biased conclusion on value, which conflicts with the valuation principles summarised above.

The aim of this paper is to show how, when using LLMs in DCF valuation, practical procedural controls tied to the valuation/measurement date can be designed so that valuation outputs remain defensible, reproducible, and auditable, and consistent with IVS and the selected basis of value. Using an illustrative case study, we demonstrate how AI-assisted estimates of a selected input variable, in this paper HICP inflation, may differ depending on whether the model is only instructed to reason as of time t , or whether it is restricted to a specific data pack available at the valuation date and required to generate an audit trail through source attribution, separation of fact from judgment, and explicit treatment of uncertainty. Based on these observations, we formulate practical recommendations for minimum controls that increase the defensibility of AI-assisted assumptions in valuation without undermining practical usability.

The paper addresses the following research question: whether simply instructing a model to reason as of the valuation date is sufficient to prevent temporal leakage and ensure consistency with the known or knowable requirement. The analysis is guided by two working propositions. First, instruction-only “as-of” reasoning does not eliminate temporal leakage. Second, closing the information window through a predefined data pack and audit trail increases methodological reliability and improves consistency with valuation-date logic.

Literature review

LLMs, as a form of generative artificial intelligence, are increasingly used in finance and valuation, both to support document work and narrative reasoning, and to form numerical assumptions for discounted cash flow (DCF)–type models (see e.g. Cheung, 2023 and 2024; Yang et al., 2024). In this context, existing studies also point to important methodological limitations.

A number of studies discuss the risk of temporal leakage (look-ahead bias), where a model unintentionally uses information that would not have been available at the relevant point in time. Building on definitions of temporal (data) leakage in time-series settings (Albelali & Ahmed, 2025), this phenomenon can be interpreted in valuation as a breach of the time boundary between the point in time as of which an estimate is formed and the information that in fact influences that estimate. Temporal leakage occurs when knowledge of subsequent developments enters a prediction or estimate, either directly (for example, through implicit use of realised values) or indirectly through reasoning shaped by facts that became known only *ex post*. As a result, the estimate may appear more accurate than it would have been under the actual conditions at that time. Zhang et al. (2026) propose detecting temporal leakage at the level of individual statements in the justification (claim-level analysis) and distinguishing between marginal contamination and decision-critical leakage. They also show that prompt instructions such as “reason as of t” may not reliably prevent contamination and that a more effective regime relies on explicit verification of claims against sources available prior to a cutoff, which in this paper corresponds to the valuation/measurement date.

Another risk relates to hallucinations, that is, situations where an LLM generates factually incorrect or source-unsupported content while remaining linguistically persuasive (Ji et al., 2023). Empirical studies (e.g. Lewis et al., 2020; Sriramanan et al., 2024) further indicate that the likelihood of such outputs increases in environments without explicit source attribution and without restricting the model to a verifiable information set. By contrast, approaches that require the model to ground its responses in specific, traceable sources reduce hallucination risk.

A further problematic aspect is the model’s sensitivity to task formulation. LLMs tend to adapt their responses implicitly to the tone, structure, and implied expectations contained in a prompt. This can lead to confirming suggested expectations (framing effect) rather than constructing assumptions through independent analytical reasoning. Even minor changes in wording can produce different results, as the model follows linguistic patterns in the input and reacts in ways that may materially affect the resulting projection (Hwang, 2026).

A specific methodological limitation also concerns how generative models handle uncertainty. LLMs can produce text that appears structured and professional, propose alternative scenarios of future developments, or offer estimated ranges. However, these outputs are not derived from an explicit economic or econometric model, so they cannot be interpreted as the result of a structural data analysis (Liu, 2024). The core issue is not only the absence of an explicit economic model but, importantly, the inability to observe which concrete information set the model relied on and how it combined individual facts. The user cannot determine which statements the model treats as verified facts, which as derived conclusions, and which as purely speculative linguistic extrapolation. This epistemic opacity means that an output may appear methodologically convincing and well structured while remaining non-reviewable in terms of its connection to information that was known or knowable as of the valuation date. In valuation, where value is normatively tied to the information set that was known or knowable as of the valuation date, this lack of transparency becomes a material obstacle to reproducibility and defensibility of AI-assisted assumptions.

Finally, potential instability across model versions should also be considered. Model updates or minor configuration changes can lead to different projections under the same task, weakening long-term reproducibility and auditability of AI-assisted estimates. The literature (see e.g. Abbasli et al., 2025) notes that this issue is not marginal but structural, as current LLMs still lack robust and stable mechanisms for estimating their own reliability.

In light of IVS requirements (see Introduction), AI prompts should include control elements that (i) close the relevant information set as of the valuation date, (ii) require explicit source references, and (iii) separate facts from judgment. These controls should also ensure that an audit trail is maintained.

Despite these recommendations, the literature so far provides limited guidance on how to transfer these control principles into valuation practice in a way that is directly compatible with IVS requirements for time specificity as of the valuation date and for restricting the relevant information set to what was known or knowable. In other words, a practically usable link between typical LLM risks and the IVS-grounded logic of the valuation date and the known or knowable information set is still missing. This paper, therefore, applies the above control principles to an AI-assisted estimate of the harmonised index of consumer prices (HICP) inflation as an input to income-based valuation using a DCF framework. By comparing different regimes of working with information, described in the next section, we show how not only the estimate itself but also its defensibility changes in terms of an IVS-compatible process.

Overall, the literature reviewed above provides the conceptual basis for the two working propositions outlined in the Introduction, which structure the empirical comparison presented in the following sections.

Methodology

The research adopts a qualitative comparative case design focused on procedural differences across three information-control regimes.

Case setting

Methodologically, the paper is based on an illustrative case study of a non-listed logistics company (Company XY) providing services to the automotive sector in the Czech Republic, including transportation, warehousing, and just-in-time or sequenced deliveries.

The relevant basis of value is the Market Value of a 100 per cent equity interest in the company for the purposes of a specified litigation context. In accordance with IVS 102 (IVSC, 2024), this basis of value reflects assumptions about typical market participants, including a hypothetical willing buyer and willing seller, and their presumed knowledge as of the valuation date 31 December 2024.

Given the nature of the subject company as a non-listed entity and the limited applicability of comparable transactions and publicly traded comparables, the primary approach applied is the income approach using a discounted cash flow entity model. Enterprise value is derived from the nominal forecast of free cash flow to the firm (FCFF) and the corresponding nominal discount rate (WACC), both determined as of the valuation date from the perspective of a hypothetical market investor. The value of 100 per cent of equity, that is, the value of the subject interest, is obtained by adjusting enterprise value for interest-bearing debt and cash as of the valuation date.

Based on on-site due diligence, revenues were identified as being driven by a combination of a volume component linked to customer production and a price component, part of which is contractually indexed to inflation. The cost base is significantly affected by wages and selected operating expenses, for which inflation also represents a key macroeconomic factor.

The forecast variable analysed in this paper is the harmonised index of consumer prices (HICP), expressed as a year-on-year inflation rate for the period following the valuation date. In the nominal DCF framework, HICP affects projected revenues and costs through inflation-related

assumptions used in nominal FCFF projections. It is therefore incorporated indirectly rather than as a separate variable in the DCF model.

This study operates with nominal projections expressed in current prices. Accordingly, the discount rate (WACC) is also determined in nominal terms as of the valuation date. The HICP forecast primarily affects (i) the price component of revenue growth and (ii) inflation-sensitive cost items. For illustrative purposes, simplified proportional relationships are assumed.

Specifically, revenue price growth is assumed to depend on HICP inflation, with the degree of inflation pass-through reflecting contractual arrangements and bargaining power. Similarly, wage growth is assumed to be linked to HICP inflation, with the extent of pass-through reflecting labour market conditions.

These simplified relationships are used only to illustrate the role of HICP within a nominal DCF framework and to show how differences in AI-assisted HICP forecasts may affect FCFF projections and, ultimately, the estimated value.

Three LLM scenarios and their purpose

The HICP forecast allows the time specificity requirement of assumptions as of the valuation date to be illustrated. The purpose of the case study is not to determine the company's value, but to compare how different regimes of AI-assisted HICP estimation, under the same valuation date, affect both the resulting estimate and its defensibility and reproducibility.

The methodological objective is to separate the predictive capability of the model from its ability to respect the valuation or measurement date and to rely exclusively on information that was known or knowable as of that date. The analysis, therefore, tests whether the instruction to reason as of time t is sufficient to prevent temporal leakage, or whether explicit control of the information set is required.

Three scenarios (A, B and C) are compared, differing in the degree of control over the information set as of the valuation date. In Scenario A (open information window, unrestricted estimate), the model is asked to reason as of the valuation date, for example by estimating expected HICP inflation for the following year from the perspective of 31 December 2024, but without closing the information set. The model is not restricted to specific materials available as of the valuation date and is not required to provide source attribution. This scenario serves as a baseline reflecting common, unrestricted LLM use and shows the estimate produced when the model may implicitly draw on a broad knowledge base, including potentially ex post information.

In Scenario B (instruction-only “as-of” setting, promised window), the model is explicitly instructed to produce the estimate as of 31 December 2024 and to rely solely on information known or knowable at that date. However, the information set is not formally closed, and no explicit data pack is provided. This scenario tests whether instruction-only “as-of” setting and declared compliance are sufficient in practice to mitigate temporal leakage and maintain a defensible valuation process without explicit input restrictions.

In Scenario C (locked estimation dataset and audit trail), the model is restricted to a predefined information set available as of the valuation date, referred to as the data pack. All factual statements and key assumptions must be supported by reference to a specific item within the data pack, identified by DocID. If no supporting evidence is available, the model is required to state insufficient evidence.

Scenario C operationalises the known or knowable requirement in a controlled environment and enables the creation of an audit trail, including prompts, model identification, source references, separation of fact from judgment, and explicit treatment of uncertainty.

The data pack is constructed to include only materials available as of 31 December 2024 and to cover core information categories relevant for HICP estimation, such as historical HICP development, macroeconomic context, and relevant institutional communication. Each document is labelled with a DocID, and the model is restricted to these items.

To minimise the risk that Scenario C results are driven by an obviously incomplete data pack, a gap check procedure is conducted. In this procedure, the model, without external browsing, assesses whether the data pack covers key information categories relevant for a 12-month HICP forecast, identifies potential missing categories, and classifies the expected impact of their absence as low, medium, or high. The gap check does not supplement the data pack with external data but serves to verify that the information set is not manifestly incomplete and that any gaps are transparently acknowledged.

Evaluation of scenario outputs

The outputs of the three scenarios are evaluated along two dimensions: numerical consistency and predictive accuracy, and procedural integrity with respect to the valuation date.

For numerical consistency and predictive accuracy, the HICP estimate is compared with ex post realised values for the selected horizon, namely the year 2025.

For procedural integrity, the analysis identifies signs of temporal leakage or look-ahead bias, the presence of unsupported statements, and the extent to which the estimate can be grounded in verifiable sources. In Scenario C, integrity is strengthened by mandatory source attribution and a documented audit trail. In Scenario B, the analysis examines whether the model can maintain internally consistent as-of reasoning without a formally closed information set.

Results and discussions

Table 1 summarises the outputs of the AI-assisted HICP year-on-year forecast for the 12 months following the valuation or measurement date of 31 December 2024 across the three information regimes. For comparison, results from two models, ChatGPT and Claude, are presented. Across scenarios and models, both forecast ranges and point estimates are relatively close. The crucial difference, however, lies in procedural defensibility in relation to the valuation date.

Table 1. Scenario-based forecasts of Czech HICP inflation for 2025 (as of 31.12.2024)

Scenario	Information window	Model	Range	Point estimate
A	Open window (unrestricted)	ChatGPT	1.7%–2.6%	2.1%
		Claude	1.6%–2.8%	2.2%
B	Promised window (instruction-only)	ChatGPT	1.8%–3.2%	2.4%
		Claude	1.8%–3.1%	2.4%
C	Closed window (locked data pack + audit trail)	ChatGPT	1.8%–3.0%	2.3%
		Claude	1.8%–3.0%	2.3%

Source: Authors' own research, based on AI-assisted forecast outputs from ChatGPT and Claude.

Scenario A represents a simple, unrestricted use of LLMs, where the model is asked to estimate inflation as of the valuation date, but without explicit restriction of the information set. ChatGPT and Claude generated point estimates of 2.1 per cent and 2.2 per cent respectively, with relatively narrow ranges between 1.6 per cent and 2.8 per cent.

The argumentative structure of both models reflected knowledge of ex post published data. ChatGPT explicitly referred to ex post information, for example stating that Czech HICP inflation slowed to approximately 2.1 per cent year-on-year by late 2025 and referring to December 2025

data. Claude used more general references such as the latest data. This contamination corresponds to temporal leakage or look-ahead bias and, at the claim level, to decision-critical leakage as discussed by Zhang et al. (2026). The models relied on ex post premises that were not known or knowable as of the valuation date. From the perspective of IVS, the outputs in Scenario A constitute a material breach of the known or knowable requirement, as the estimate is influenced by information not available at the valuation date.

In Scenario B, we explicitly instructed the model to rely only on information known or knowable as of 31 December 2024. Unlike Scenario A, the model worked with 2024 information, acknowledged the potential absence of December data, and explicitly stated uncertainties. The point estimate was 2.4 per cent, with a wider range of 1.8 per cent to 3.2 per cent, which may indicate the absence of newer data or ex post knowledge of realised trends.

The output appears temporally consistent, and ChatGPT explicitly recognised uncertainty about data availability. However, Scenario B suffers from a different weakness. The information set remains formally open, and the model is not required to link statements to identifiable sources. Without a closed information window and without an audit trail, it is not possible to verify whether the premises used were actually available as of the valuation date or whether they reflect generalised patterns embedded in the model.

Thus, although the model is instructed to rely only on information known or knowable at the valuation date, the user cannot verify whether this requirement is really met. From an IVS perspective, this lack of verifiability is insufficient. The limitation of Scenario B, therefore, lies less in explicit temporal leakage and more in the non-reviewability of the origin of premises and the inability to separate facts from judgment. In addition, without mandatory source attribution, the risk of hallucinations remains present, even where the output appears “as-of” consistent. The residual epistemic opacity remains unresolved.

In Scenario C, the model was restricted to a closed information set consisting of three specific documents available as of 31 December 2024: the European Commission Autumn 2024 Economic Forecast, the IMF World Economic Outlook October 2024, and the OECD Economic Outlook November 2024. The prompt explicitly required that each factual statement be supported by reference to a specific DocID and that the model state insufficient evidence where no support was available. Although technical browsing capability was available, the protocol defined the data pack as the sole permissible information source, and the model was instructed not to use or reference any external materials.

This regime substantially reduces the scope for hallucinations in the form of unsupported factual claims, as no factual statement can be accepted without linkage to a document in the data pack. It does not eliminate judgment beyond the evidence, but such judgment is explicitly identified as such. Temporal leakage is significantly reduced at the level of factual premises because the information window is formally closed in line with IVS requirements.

As part of Scenario C, a gap check was conducted to assess whether the data pack covered standard categories of information typically relevant for a 12-month HICP forecast. The check showed that the data pack contained general macroeconomic information relevant for inflation developments in 2025. However, it did not include monthly Czech HICP or CPI statistics for 2024, detailed projections of Czech inflation, component breakdowns, or explicit monetary policy assumptions that would allow a mechanical derivation of a precise point estimate. The model correctly identified these limitations as insufficient evidence.

The resulting point estimate of 2.3 per cent and the range of 1.8 per cent to 3.0 per cent therefore represent professional judgment applied within an evidence-based framework of near-

target inflation. The wider interval reflects the limited granularity of the data pack and the necessary reliance on judgment.

Compared with Scenarios A and B, Scenario C enables transparent review of each factual statement and verification of its known or knowable status as of the valuation date. The risk of temporal leakage is substantially reduced at the factual level. Nevertheless, residual epistemic opacity remains, as even with explicit DocID references, it is not observable how the model weights individual documents when forming its numerical conclusion.

Repeated generations under identical prompts produced minor variations in wording, point estimates, and interval widths, even when the prompt was entered by different users within the same service environment. This confirms that outputs are not fully deterministic and may vary across replications, consistent with literature highlighting potential instability of generative models across versions and configurations (Abbasli et al., 2025). For reproducibility, it is therefore necessary to archive at least the model identification, the exact prompt wording, and the data pack used. Among the tested regimes, Scenario C provides the most defensible output in terms of IVS compatibility, as it establishes an auditable chain of factual premises, respects the valuation-date boundary, and clearly separates evidence-based premises from judgment-based assumptions.

For the purpose of assessing numerical consistency and predictive accuracy, the AI-assisted estimates were compared with ex post realised HICP values for 2025. According to official Eurostat data, the Czech Republic recorded an average annual HICP rate of 2.3 per cent in 2025 (Eurostat, 2026). This realised value coincides exactly with the 2.3 per cent point estimate produced by both ChatGPT and Claude in Scenario C and is close to the estimates in Scenarios A and B.

All three scenarios, therefore, appear numerically consistent with the subsequently realised inflation outcome. From a predictive perspective, the LLM estimates cluster closely around the actual value. However, numerical proximity alone is insufficient for IVS-compliant valuation use. What is decisive is whether assumptions are derived exclusively from information known or knowable as of the valuation date and whether their origin is auditable.

In this experiment, the closest numerical match occurred in Scenario C, the regime with the strictest information controls. This result cannot be generalised, but it is worth noting. It suggests that strict restriction of the information set does not necessarily degrade forecast quality. By contrast, the numerical accuracy observed in Scenarios A and B must be interpreted with caution because those regimes did not ensure effective control over the information set as of the valuation date. Taken together, the findings support the two working propositions formulated in the Introduction. Proposition 1 is supported by the finding that simply instructing the model to reason as of the valuation date may reduce obvious use of ex post information. However, it does not eliminate epistemic opacity or ensure verifiable compliance with the known or knowable requirement. Proposition 2 is supported by the results from Scenario C. Closing the information window through a predefined data pack and mandatory source references improved transparency, reviewability, and consistency with IVS requirements.

Conclusion

Our findings show that AI can be useful in preparing input assumptions for income-based valuation, particularly by helping to organise information and formulate assumptions more consistently. However, the paper also reveals important methodological challenges and risks.

An AI-assisted estimate may be close to the value that is later realised. However, under IVS, the estimate may become problematic if it relies on information that was not known or knowable at the valuation date. Valuation should be based only on information available at that

time and the process should be transparent. Otherwise, the use of later information may create look-ahead bias and reduce the credibility of the valuation result.

The comparison of the three scenarios suggests that simply telling the model to reason “as of” the valuation date (Scenario B) is not sufficient. Although the outputs may appear time-consistent, the user still cannot verify whether the estimate relies only on information that was available at that time. In Scenario A, where the information window remains open, the model may also be influenced by ex post information. However, this may not necessarily be obvious from the output itself. Such estimates may appear persuasive and numerically accurate, but they are still in conflict with IVS requirements.

A key finding is that higher procedural discipline and a closed information set (Scenario C) produced results that were consistent not only with IVS requirements but, in this experiment, also with subsequently realised values. This suggests that closing the information window does not necessarily reduce forecast quality. At the same time, it makes the assumptions more transparent and easier to defend. Nevertheless, these findings cannot be generalised without further testing.

From a practical perspective, the paper suggests a minimum set of controls that should be used when AI-assisted valuation is applied. First of all, it is necessary to define a closed information window as of the valuation date. Another requirement is to use a predefined data pack and to provide source references for factual statements. The process should also explicitly acknowledge insufficient evidence, separate facts from professional judgment, and maintain an audit trail containing the model version, prompts, data pack, and key expert decisions. These controls are not only good practice when working with AI, but also help ensure that the valuation process remains consistent with IVS requirements regarding the valuation date, transparency, and defensibility of assumptions. At the same time, important limitations remain. Even in Scenario C with a closed information window and explicit DocID references, the user still cannot fully observe how the model combines individual facts when forming the final estimate. Some degree of epistemic opacity therefore remains. In addition, outputs are not fully deterministic and may vary across repeated runs of the same task, which highlights the importance of documentation and audit trails.

Finally, the study is based on one illustrative case and a single DCF input variable, so the findings should be interpreted with caution. Future research should examine whether the proposed controls also work for other DCF inputs, such as revenue growth, margins, or discount rate assumptions, and in different valuation contexts. It would also be useful to test how sensitive the outputs are to changes in prompts and model settings.

The main contribution of this paper lies in linking common LLM risks, such as temporal leakage, hallucinations, and epistemic opacity, to the application of IVS requirements in valuation. Previous studies discuss these risks mainly in general forecasting and decision-making contexts. This paper shows that the same risks may also affect AI-assisted DCF valuation. The findings also suggest that controlling the information set and maintaining an audit trail can improve the transparency and defensibility of the valuation process under IVS.

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT by OpenAI and Grammarly to assist with language editing and improving readability. All technical content, interpretations, and conclusions remain the responsibility of the authors, who take full responsibility for the final content of the publication.

References

- Abbasli, T., Toyoda, K., Wang, Y., Witt, L., Ali, M. A., Miao, Y., Li, D., & Wei, Q. (2025). Comparing uncertainty measurement and mitigation methods for large language models: A systematic review. <https://arxiv.org/abs/2504.18346>
- Albelali, S., & Ahmed, M. (2025). Hidden leaks in time series forecasting: How data leakage affects LSTM evaluation across configurations and validation strategies. <https://arxiv.org/abs/2512.06932>
- Cheung, K. S. (2023). Real estate insights: Unleashing the potential of ChatGPT in property valuation reports: The “Red Book” compliance chain-of-thought (CoT) prompt engineering. *Journal of Property Investment & Finance*, 42(2), 200–206. <https://doi.org/10.1108/JPIF-06-2023-0053>
- Cheung, K. S. (2024). Real estate insights: Establishing transparency—Setting AI standards in property valuation. *Journal of Property Investment & Finance*, 42(4), 406–408. <https://doi.org/10.1108/JPIF-04-2024-0050>
- Eurostat. (2026). *Harmonised index of consumer prices (HICP) – ECOICOP ver. 2 – indices and rates of change, annual data* [Data set]. https://ec.europa.eu/eurostat/databrowser/view/prc_hicp_ainr/default/table?lang=en
- Hwang, Y., Lee, D., Kang, T., Lee, M., & Jung, K. (2026). When wording steers the evaluation: Framing bias in LLM. <https://arxiv.org/abs/2601.13537>
- International Valuation Standards Council. (2024). *International Valuation Standards* (effective 31 January 2025). <https://ivsc.org/standards/>
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), Article 248, 1–38. <https://doi.org/10.1145/3571730>
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems* (Vol. 33). <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>
- Liu, L., Pan, Y., Li, X., & Chen, G. (2024). Uncertainty estimation and quantification for LLMs: A simple supervised approach. <https://arxiv.org/abs/2404.15993>
- Sriramanan, G., Bharti, S., Sadasivan, V. S., Saha, S., Kattakinda, P., & Feizi, S. (2024). LLM-Check: Investigating detection of hallucinations in large language models. In *Advances in Neural Information Processing Systems* (Vol. 37). https://proceedings.neurips.cc/paper_files/paper/2024/hash/3c1e1fdf305195cd620c118aaa9717ad-Abstract-Conference.html
- Yang, X., Zang, S., Ren, Y., Peng, D., & Wen, Z. (2024). Evaluating large language models on financial report summarization: An empirical study. <https://arxiv.org/abs/2411.06852>
- Zhang, Z., Chen, R., & Stadie, B. C. (2026). All leaks count, some count more: Interpretable temporal contamination detection in LLM backtesting. <https://arxiv.org/abs/2602.17234>