# LLM Accuracy in Tabular Fact Verification: A Factorial Evaluation of Model, Temperature, and Prompting Strategy

Rafael Camacho-Aguilar, Javier Rojas-Segura* and José Martínez-Villavicencio

*Instituto Tecnológico de Costa Rica, Cartago, Costa Rica*

**Abstract**

Small and medium-sized enterprises are increasingly adopting large language models for analytic verification tasks without the infrastructure that larger organizations rely on to validate model outputs. This study evaluates three commercially available large language models (Gemini 3.1 Flash Lite, DeepSeek-V3, and Llama 3.3 70B Turbo) on the TabFact benchmark using a full factorial design (3 models $\times$ 3 temperatures $\times$ 2 prompting strategies $\times$ 5 iterations, totalling 9,000 inferences). Accuracy ranges from 83.5% to 86.7%, systematically below the human benchmark of 92.1% [1]; Zero-Shot prompting equals or outperforms Chain-of-Thought across all models ($\Delta = +1.4$ pp, Wilcoxon $p = 0.038$); temperature has no significant effect ($H = 0.17$, $p = 0.921$); and models exhibit severe overconfidence (Expected Calibration Error of 0.136). Reproducing the full factorial experiment costs USD $2.45 in total across the three models (USD $0.28–$1.47 per model) under current public list rates, bringing evidence-based model selection within reach of resource-constrained organizations. The most expensive model costs $5.35\times$ what the cheapest does for an additional 3.2 pp of accuracy, indicating that model selection should be driven by the downstream cost of incorrect verification rather than by headline benchmark performance.

**Keywords**

Large Language Models (LLM), Tabular Fact Verification, TabFact, Prompt Engineering, Calibration, Cost-Effectiveness, SME Governance

## 1. Introduction

The World Economic Forum reports that 86% of employers expect artificial intelligence (AI) and information-processing technologies to transform their business by 2030 [2]; the same survey identifies skills gaps as the principal barrier to that transformation, with 63% of respondents citing them as the foremost obstacle to organizational change. This asymmetry between expected transformation and recognized capacity gap is not uniformly distributed: large corporations with dedicated machine-learning operations (MLOps) teams can validate what they deploy [3]; small and medium-sized enterprises (SME) in emerging economies, operating with constrained budgets and without in-house data infrastructure, cannot [4]. In Latin America, where commercial AI adoption is accelerating alongside a documented shortage of technical talent [5], the governance question is not whether large language models (LLM) will enter reporting and verification workflows, since that adoption is already underway, but whether the organizations deploying them can detect when their outputs are wrong.

Tabular fact verification, determining whether a natural language statement is consistent with a structured data table, carries practical scope well beyond an academic benchmark: it is the precise task executed when a manager asks a language model whether a dashboard figure is correct, when an auditor queries whether a report reconciles with its underlying data source, or when a compliance system checks whether a regulatory indicator was computed according to its definition [1]. The skill

*IWSAI 2026: 1st International Workshop on Systems, Applications, Artificial Intelligence and Innovation, September 16–18, 2026, Piura, Peru*

*Corresponding author.

fitacam@gmail.com (R. Camacho-Aguilar); jarojas@tec.ac.cr (J. Rojas-Segura); jomartinez@tec.ac.cr (J. Martínez-Villavicencio)

0009-0005-6562-509X (R. Camacho-Aguilar); 0000-0002-0488-4056 (J. Rojas-Segura); 0000-0002-7576-4625 (J. Martínez-Villavicencio)

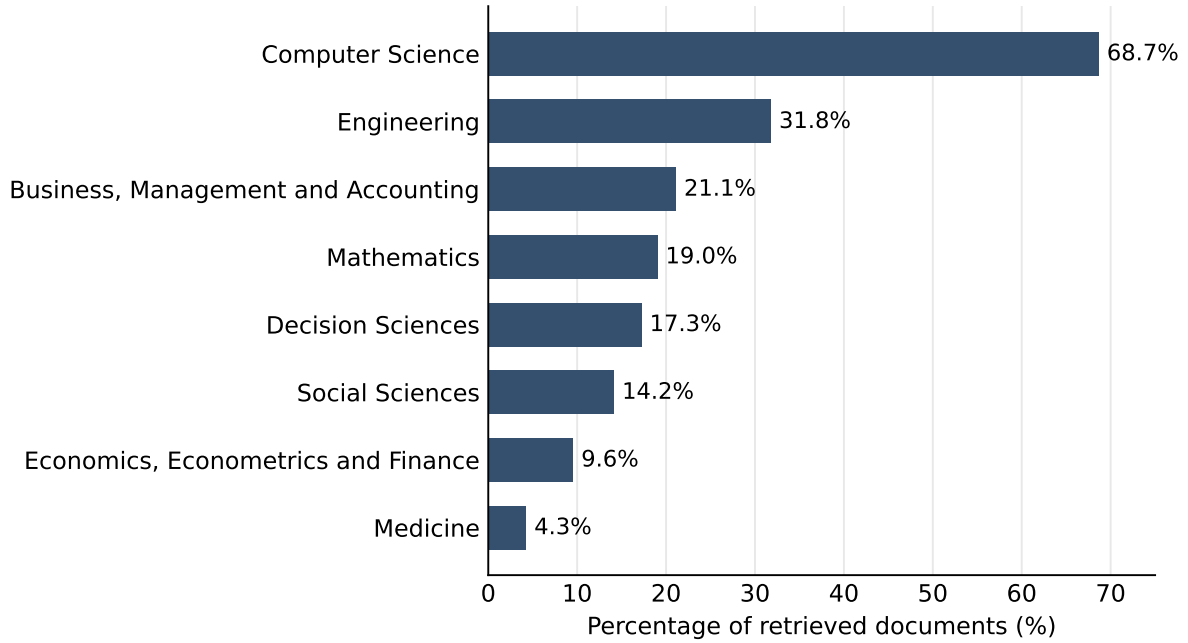


**Figure 1:** Disciplinary distribution of the retrieved literature on LLMs in enterprise and SME contexts (Scopus; $N = 2{,}742$; 2021–2026; leading ASJC subject areas). Percentages exceed 100% because Scopus assigns documents to multiple subject areas.

combination required (linguistic attribute matching, negation handling, and symbolic operations such as counting, ranking, and aggregation) makes tabular fact verification representative of the analytic verification workload that organizations are beginning to delegate to LLMs.

Two gaps motivate this study. The first is methodological: existing evaluations of LLMs on TabFact examine single models or fixed prompting strategies [6, 7], without a factorial design that simultaneously varies model, temperature, and prompting strategy against a published human baseline. The second is disciplinary: a focused bibliometric search on LLM literature published between 2021 and 2026 that explicitly engages with enterprise and SME contexts reveals that 68.7% of this output appears in Computer Science venues, while only 21.1% comes from Business, Management and Accounting fields[1]. Figure 1 shows the broader pattern: the technical disciplines, led by computer science and followed by engineering and mathematics, dominate the literature, while the business and social-science fields closest to organizational deployment remain comparatively underrepresented. The result leaves organizational decision-makers with evidence calibrated predominantly to technical rather than operational perspectives.

This paper addresses both gaps through a full factorial evaluation of three commercially available LLMs on the TabFact benchmark [1], guided by a single question: how reliably do these models verify tabular claims, and can a resource-constrained organization tell when they are wrong?

## 2. Background and Related Work

### 2.1. LLMs in Structured Reasoning Tasks

Verifying facts against structured tables is a core challenge in natural language processing. The TabFact benchmark [1] has emerged as a standard reference for this task, pairing 118,000 human-

[1]Figures derived from a Scopus search (performed 25 May 2026) using the query `TITLE-ABS-KEY (("large language model*" OR "generative AI" OR "generative artificial intelligence") AND ("enterprise*" OR "SME*" OR "small and medium" OR "firm*"))` without document-type restriction, returning $N = 2{,}742$ documents (2021–2026). Percentages refer to All Science Journal Classification (ASJC) subject areas; categories total above 100% because Scopus assigns documents to multiple ASJC areas.

annotated statements with 16,000 Wikipedia tables. Early neural baselines, such as Table-BERT and the Latent Program Algorithm, achieved roughly 68% accuracy, falling well short of the 92.1% human performance ceiling. Subsequent techniques that pre-trained models specifically on tabular data closed much of this gap, with TAPAS [6] reaching 81% and TAPEX [7] hitting 84% on the test set. More recently, the arrival of general-purpose LLMs [8] has shifted the paradigm, making zero- and few-shot verification possible without task-specific training. However, existing studies rarely evaluate the combined impact of temperature settings and prompting strategies, leaving their joint effect on model reliability largely unmeasured; our research introduces a factorial design to systematically test both variables simultaneously.

These reliability concerns connect to a wider debate about the nature of LLMs themselves. The literature offers two opposing characterizations of these models: Bender et al. [9] describe them as "stochastic parrots" that produce plausible text without semantic grounding, whereas Park et al. [10] present them as "generative agents" capable of simulating human-like decision behavior. Recent surveys document the rapid uptake of LLMs in computational social science applications [11], where reliability under operational conditions becomes critical. By measuring how consistently these models verify structured claims, our study offers evidence on which framing better describes their behavior in analytic verification.

## 2.2. Prompt Engineering: Zero-Shot vs. Chain-of-Thought

Chain-of-Thought (CoT) prompting [12] asks a language model to lay out its reasoning step by step before committing to an answer, and it substantially improves performance on arithmetic and commonsense reasoning benchmarks. Zero-Shot-CoT [13] showed that the technique works even without worked examples in the prompt. Because individual reasoning chains can be inconsistent, later work sampled and aggregated several of them to stabilize the output [14]. The gains, however, concentrate on complex, multi-step problems [12]: when an answer can be read directly off structured data, forcing the model to reason aloud can add steps that introduce errors rather than prevent them. Tabular fact verification is plausibly one of these cases, since the evidence already sits in discrete cells and checking a claim is usually a direct lookup or comparison.

## 2.3. Calibration and Confidence in LLMs

Guo et al. [15] established Expected Calibration Error (ECE) as a standard metric for assessing whether a model's confidence estimates align with its empirical accuracy. Models that report high confidence regardless of correctness are particularly problematic in high-stakes contexts, because they deprive decision-makers of meaningful uncertainty signals. Recent work nonetheless reports that large language models can, to varying degrees, assess the reliability of their own answers [16], a self-knowledge that proves uneven across tasks and output formats. Bhatt et al. [17] argued that well-calibrated uncertainty serves as a form of transparency in AI systems, giving decision-makers actionable signals about where to allocate human scrutiny. Without it, the theoretical risks of language models become operational harms [18]. In an organizational setting, verification outputs that are wrong yet highly confident can propagate silently into downstream decisions without ever triggering review.

## 2.4. LLMs in SME and Resource-Constrained Settings

The adoption of generative AI tools by SMEs has accelerated considerably, particularly in regions where consumer-ready tools have lowered traditional adoption barriers [5]; yet the literature on how these organizations validate the outputs they receive remains comparatively sparse. Existing technical evaluations implicitly assume enterprise contexts with MLOps infrastructure, dedicated data science teams, and internal validation pipelines [3]; these assumptions do not hold for an SME that adopts an LLM as an external application programming interface (API) service and lacks the technical capacity to design evaluation pipelines of its own. The operational reliability concerns documented in the LLM

literature thus apply with particular force in these settings, where there is no internal mechanism to catch errors before they propagate into decisions.

While current research races to push capability frontiers and improve computational efficiency, it often overlooks the practical needs of resource-constrained organizations. Managers lack the actionable evidence they need to answer basic questions. When is it safe to let a model verify a claim? Which of its signals can they trust? Which configurations reduce deployment risk? Without an empirical baseline suited to their conditions, decision-makers adopt powerful technology blind. The operational harms of LLMs [18] threaten organizations of every size, which makes the transparency that well-calibrated uncertainty provides [17] essential for non-expert users. Our study addresses exactly this gap. By crossing models, temperatures, and prompting strategies on a structured verification task with commercially available APIs, our factorial design produces a reference baseline that organizations can interpret and replicate without specialized infrastructure.

# 3. Method

## 3.1. Experimental Design

We adopt a full factorial design with four factors: model (Gemini 3.1 Flash Lite, DeepSeek-V3, Llama 3.3 70B Turbo), temperature ($T \in \{0.0, 0.5, 1.0\}$), prompting strategy (Zero-Shot, Chain-of-Thought Expert), and iteration (5 repetitions per cell). This yields $3 \times 3 \times 2 \times 5 = 90$ experimental cells applied to 100 TabFact cases, producing 9,000 inferences in total.

## 3.2. Dataset

We sample 100 cases from the TabFact test split,[2] which comprises 12,779 statements total [1], using stratified sampling with seed $= 42$. The sample is balanced across two dimensions simultaneously: ground-truth label (50 ENTAILED + 50 REFUTED) and reasoning surface form (50 linguistic + 50 numerical), yielding exactly 25 cases per design cell (Table 1). Reasoning type was classified by a regex-based heuristic detecting numeric patterns in the statement text; this classification is independent of TabFact's original simple/complex channel split.

**Table 1**
Sample design: cases per cell ($N = 100$)

| | Linguistic | Numerical |
|---|---|---|
| Entailed | 25 | 25 |
| Refuted | 25 | 25 |

Tables in the sample contain between 5 and 42 rows (mean $= 14.7$, SD $= 9.0$), consistent with TabFact's design criterion of fewer than 50 rows per table. Mean table size is comparable across labels (Entailed: $\bar{x} = 15.3$; Refuted: $\bar{x} = 14.2$), and mean statement length is similar across reasoning types (linguistic: 12.9 words; numerical: 14.0 words), confirming that the two stratification dimensions do not introduce systematic complexity confounds. All 100 cases are unique (verified by content hash).

**Illustrative cases.** Table 2 shows two cases from the sample, one per label, with statements reproduced verbatim. Case (a) is a linguistic REFUTED statement: verification requires locating the minimum of the *crowd* column and checking its venue — a single lookup, phrased in TabFact's characteristically telegraphic grammar. Case (b) is a numerical ENTAILED statement: verification requires aggregating medal counts per nation, ranking them, and resolving a tie for second place. The pair spans the skill range of the task, from direct attribute matching to multi-step symbolic operations over the table.

[2] HuggingFace dataset: `table-benchmark/tabfact`.

**Table 2**
Two illustrative cases from the evaluation sample (statements reproduced verbatim)

**(a) Linguistic, gold label REFUTED:** *"the smallest crowd be at venue mcg"*

| home team | venue | crowd |
|---|---|---|
| geelong | kardinia park | 25723 |
| fitzroy | brunswick street oval | 23012 |
| south melbourne | lake oval | 14350 |
| melbourne | mcg | 31455 |
| north melbourne | arden street oval | 15000 |
| footscray | western oval | 21639 |

Table "1961 VFL season"; three of seven columns shown.

**(b) Numerical, gold label ENTAILED:** *"the nation with the most medal be italy with 3 and the nation with second most medal be west germany and austria with 2"*

| rank | nation | gold | silver | bronze | total |
|---|---|---|---|---|---|
| 1 | italy | 2 | 1 | 0 | 3 |
| 2 | west germany | 1 | 1 | 0 | 2 |
| 3 | austria | 0 | 1 | 1 | 2 |
| 4 | poland | 0 | 0 | 1 | 1 |
| 5 | east germany | 0 | 0 | 1 | 1 |

Table "FIL World Luge Championships 1971"; full table.

## 3.3. Models and API Configuration

Three commercially available LLMs were evaluated via their public APIs. Gemini 3.1 Flash Lite (Google) is a lightweight, high-throughput model in the Gemini 3 series. DeepSeek-V3 (DeepSeek AI) is an open-weight model accessed via the DeepSeek API. An important configuration note: the DeepSeek API applies an internal temperature remapping ($T_{\text{model}} = T_{\text{api}} \times 0.3$ for $0 \leq T_{\text{api}} \leq 1$), as documented on the official model card [19], so effective temperatures are not strictly comparable across models. Llama 3.3 70B Instruct Turbo (Meta AI) is an open-weight instruction-tuned model accessed via Together.ai.[3] All models returned JavaScript Object Notation (JSON) responses containing a binary verdict (ENTAILED/REFUTED), a confidence score (1–10), and a brief reasoning trace.

## 3.4. Evaluation Metrics

- *Accuracy:* proportion of correctly classified statements (95% Wilson confidence intervals).
- *Format Failure Rate (FFR):* proportion of invalid JSON responses excluded from analysis.
- *Expected Calibration Error (ECE):* area-weighted deviation between confidence and accuracy across 10 bins.
- *Fleiss' $\kappa$:* inter-model agreement on binary verdicts [20].
- *Cohen's $h$:* effect size for proportion comparisons [21].

**Statistical testing.** For pairwise model accuracy comparisons we apply chi-square tests with Bonferroni correction ($\alpha = 0.05$); Wilson 95% confidence intervals are reported for all proportions in preference to normal approximation intervals, which are unreliable near boundary values [22]. For the prompting strategy comparison we use the Wilcoxon signed-rank test on case-level accuracy averaged across temperatures and iterations ($n = 100$ pairs), which accounts for the paired structure of the design. For temperature effects we apply Kruskal-Wallis $H$, appropriate for bounded proportion data. Effect

[3]API model identifiers used: `gemini-3.1-flash-lite` (Google AI Studio), `deepseek-chat` (DeepSeek API), and `meta-llama/Llama-3.3-70B-Instruct-Turbo` (Together.ai).

sizes are reported as Cohen's $h$ for proportions and Cohen's $d$ for continuous variables (confidence scores) [21].

### 3.5. Cost Estimation Methodology

Because the experiment was executed without per-inference token logging, operational costs were estimated post hoc rather than measured. For each valid inference ($n = 8{,}986$ of $9{,}000$; see Section 4.1), input tokens were estimated from the prompt text (system instruction, table serialization, and statement) and output tokens from the complete API response captured in the experiment records. Token counts were obtained by a character-length heuristic (tokens $\approx$ chars$/4$), the standard approximation used when a model-specific tokenizer is unavailable. This heuristic introduces an aliasing error of roughly $\pm 10$–$15\%$ relative to provider-specific tokenizers; absolute costs reported should therefore be interpreted as order-of-magnitude estimates rather than as billing-grade figures. Two features nonetheless limit how much this approximation affects the comparative findings. The same heuristic is applied identically to all three models, so any systematic bias in the character-to-token conversion partially cancels when costs are expressed as cross-model ratios; moreover, the per-token rates are exact provider-published figures rather than estimates. The cost ratios on which the deployment argument rests are accordingly far less sensitive to the estimation error than the absolute figures.

Per-token rates were taken from each provider's official public documentation, retrieved on May 14, 2026 (Table 3). For DeepSeek-V3 the cache-miss input rate was applied throughout, since the experiment did not exploit prompt caching. The total estimated cost across the three models is reported in Section 4; the absolute figure is of limited interest here, since experimental costs at this scale are small; what matters are the cross-model ratios, which directly inform deployment decisions for organizations that must select among commercially available LLMs under budget constraints.

**Table 3**
Public list rates per million tokens (USD), retrieved May 14, 2026.

| Model (provider) | Input | Output |
|---|---|---|
| Gemini 3.1 Flash Lite (Google) | 0.25 | 1.50 |
| DeepSeek-V3 (DeepSeek) | 0.14 | 0.28 |
| Llama 3.3 70B Turbo (Together) | 0.88 | 0.88 |

## 4. Results

### 4.1. Data Quality

Of 9,000 inferences, 8,986 (99.84%) produced valid JSON responses (FFR $= 0.16\%$). DeepSeek-V3 accounted for most failures (FFR $= 0.37\%$, concentrated in the Zero-Shot condition), while Gemini 3.1 Flash Lite and Llama 3.3 70B Turbo showed near-zero failure rates ($\leq 0.07\%$). All subsequent analyses use the 8,986 valid inferences.

### 4.2. Overall Accuracy

Table 4 reports accuracy per model against the human benchmark. All three LLMs fall short of the 92.1% human reference [1], with a gap of 5.4 to 8.6 percentage points. Pairwise chi-square tests with Bonferroni correction show that only the Llama–DeepSeek difference reaches significance ($\chi^2(1) = 12.24$, $p = 0.001$, $h = 0.09$); the remaining pairs do not ($p > 0.08$).

### 4.3. Effect of Prompting Strategy (CoT Paradox)

Zero-Shot equals or outperforms CoT across all models (Figure 2; $\Delta = +1.4$ pp; Wilcoxon signed-rank, $W = 214.5$, $p = 0.038$; positive $\Delta$ indicates Zero-Shot accuracy exceeds CoT accuracy). By model:

**Table 4**
Model accuracy on TabFact (8,986 valid inferences, 100 cases)

| Model | Accuracy | 95% CI |
|---|---|---|
| Human [1] | 92.1% | — |
| Llama 3.3 70B Turbo | 86.7% | [85.5%, 87.9%] |
| Gemini 3.1 Flash Lite | 85.6% | [84.3%, 86.8%] |
| DeepSeek-V3 | 83.5% | [82.1%, 84.8%] |

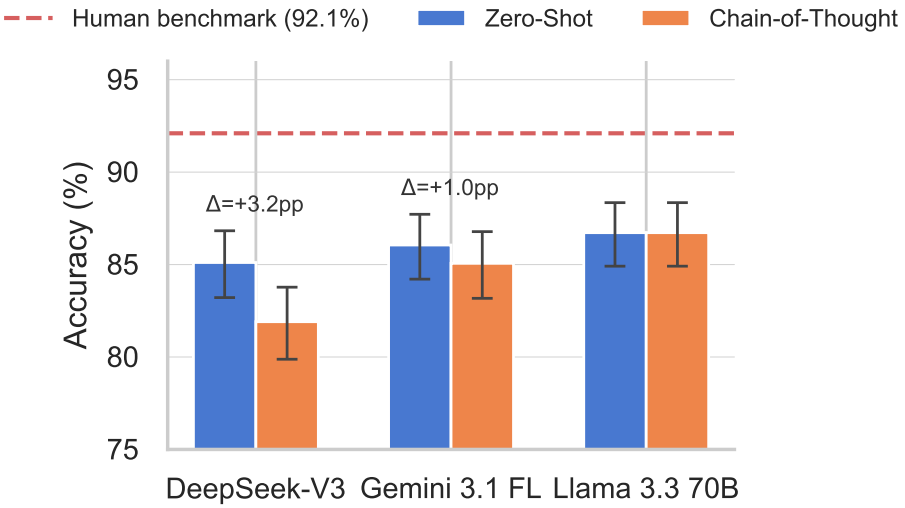


**Figure 2:** Accuracy by model and prompting strategy on TabFact (8,986 valid inferences). Error bars: Wilson 95% confidence intervals. The dashed line marks the human benchmark (92.1%; [1]). Zero-Shot consistently equals or outperforms Chain-of-Thought across all models; the gap is largest for DeepSeek-V3 ($\Delta = +3.2$ pp) and vanishes for Llama 3.3 70B Turbo ($\Delta = 0.0$ pp).

DeepSeek-V3 shows the largest gap in favor of Zero-Shot (ZS $= 85.1\%$ vs. CoT $= 81.9\%$, $\Delta = +3.2$ pp), Gemini 3.1 Flash Lite shows a smaller advantage (ZS $= 86.1\%$ vs. CoT $= 85.1\%$, $\Delta = +1.0$ pp), and Llama 3.3 70B Turbo is completely indifferent (ZS $=$ CoT $= 86.7\%$, $\Delta = 0.0$ pp). This counterintuitive result suggests that explicit step-by-step reasoning introduces noise rather than precision in structured tabular verification.

## 4.4. Effect of Temperature

Temperature had no significant effect on accuracy (Kruskal-Wallis $H = 0.17$, $p = 0.921$; range: 85.1–85.5%), indicating that for structured binary classification the stochastic sampling component adds no systematic bias. The null applies to temperature as set through each provider's API: because DeepSeek internally rescales temperature, the three levels are not strictly comparable across models, so the absence of an effect is best read at the level of the API controls that organizations actually use. This result is practically useful: temperature is not a critical tuning parameter for this task type.

## 4.5. ENTAILED vs. REFUTED Accuracy Gap

Models classified REFUTED statements more accurately than ENTAILED ones (89.1% vs. 81.4%; Cohen's $h = 0.22$, $p = 8.5 \times 10^{-25}$), revealing a systematic verification asymmetry. The gap varies substantially across models (Table 5): DeepSeek-V3 exhibits the largest asymmetry (15.3 pp), while Llama 3.3 70B Turbo is most balanced (2.4 pp).

**Table 5**
Accuracy by label type and model

| Model | ENTAILED | REFUTED | Gap |
|---|---|---|---|
| DeepSeek-V3 | 75.8% | 91.1% | 15.3 pp |
| Gemini 3.1 Flash Lite | 82.9% | 88.3% | 5.4 pp |
| Llama 3.3 70B Turbo | 85.5% | 87.9% | 2.4 pp |

## 4.6. Error Patterns by Table Type

Aggregate error rates conceal a strongly concentrated error structure: 59 of the 100 cases were classified correctly in every valid inference, while the ten most error-prone cases account for 54.8% of all 1,324 errors (the twenty most error-prone, for 88.4%). Two distinct error patterns emerge.

The first pattern is mechanical and shared across models: table length (Table 6). Statements over large tables (21–42 rows) were misclassified in 28.2% of inferences, against 12.8% for small (5–10 rows) and 8.6% for medium tables (11–20 rows); the effect replicates in each model separately and within both reasoning types (25.6% linguistic, 30.7% numerical). Reasoning surface form itself, by contrast, is not a shared difficulty axis: pooled error rates are nearly identical for linguistic and numerical statements (14.8% vs. 14.7%), and the direction of the difference is model-specific (DeepSeek-V3 errs more on numerical statements, 18.7% vs. 14.3%, while Gemini 3.1 Flash Lite and Llama 3.3 70B Turbo err more on linguistic ones). The label asymmetry documented above also interacts with reasoning type: ENTAILED-numerical statements form the hardest cell (19.9% error) and REFUTED-numerical the easiest (9.5%), consistent with an asymmetric verification burden — confirming a numerical claim requires checking every relevant cell, whereas refuting one requires a single counterexample.

The second pattern is semantic and case-specific. Only six cases were misclassified by all three models in a majority of their inferences, and just three exceeded 90% error; inspection shows statements whose truth conditions are ambiguous relative to the table (for instance, whether a claim of the form "engine X has chassis Y" asserts co-occurrence or exclusivity), on which the three models converge on the same verdict. Two of these cases fall in the small-table bin, where they alone contribute 35% of its errors; excluding the three ambiguous cases, error rates are flat at 8.6–8.7% for tables of up to 20 rows and rise to 25.0% beyond, in all three models. Misclassification is therefore not diffuse noise: it concentrates where evidence retrieval spans long tables or where the statement itself underdetermines the verdict — two conditions an organization can screen for before delegating verification to an LLM.

**Table 6**
Error rate (%) by table size and model (8,986 valid inferences)

| Model | 5–10 rows | 11–20 rows | 21–42 rows | All |
|---|---|---|---|---|
| DeepSeek-V3 | 14.3 | 12.4 | 27.3 | 16.5 |
| Gemini 3.1 Flash Lite | 14.8 | 6.3 | 26.4 | 14.4 |
| Llama 3.3 70B Turbo | 9.2 | 7.3 | 30.8 | 13.3 |
| Pooled | 12.8 | 8.6 | 28.2 | 14.7 |

## 4.7. Confidence Calibration

Models reported maximum confidence (score $= 10$) on 89.2% of responses, yielding ECE $= 0.136$. The mean confidence delta between correct and incorrect predictions was only 0.09 points (scale 1–10), meaning a manager cannot use a model's self-reported certainty to judge whether an answer is correct. Model-level analysis reveals heterogeneity: Llama 3.3 70B Turbo shows the strongest, though still modest, discrimination (Cohen's $d = 0.46$, $p < 0.001$), DeepSeek-V3 shows moderate discrimination ($d = 0.33$, $p < 0.001$), while Gemini 3.1 Flash Lite shows no statistically significant discrimination ($d = 0.17$, $p = 0.18$).

### 4.8. Inter-Model Agreement

Fleiss' $\kappa = 0.740$ [20] indicates substantial overall agreement on verdict labels, yet models diverge on 1 in every 5 evaluation groups, precisely the edge cases most relevant to auditability. Agreement decreases at higher temperatures ($\kappa = 0.712$ at $T = 1.0$ vs. $\kappa = 0.760$ at $T = 0.0$), consistent with increased output stochasticity.

### 4.9. Cost Analysis

The total estimated cost for the 8,986 valid inferences across the three models was USD $2.45, distributed as shown in Table 7. DeepSeek-V3 was the cheapest model by a substantial margin (USD $0.28 total), Gemini 3.1 Flash Lite occupied the middle position (USD $0.70), and Llama 3.3 70B Turbo was the most expensive (USD $1.47). The cost ratios diverge sharply from the accuracy ratios (Figure 3): Llama costs $5.35\times$ what DeepSeek does while delivering only 3.2 pp of additional accuracy, a trade-off whose attractiveness depends on the downstream cost of an incorrect verification.

**Table 7**
Estimated API cost per model ($n = 8{,}986$ valid inferences). Rates from Table 3.

| Model | Total (USD) | Mean/inf. (USD) | vs. cheapest |
|---|---|---|---|
| DeepSeek-V3 | 0.275 | 0.000092 | 1.00× |
| Gemini 3.1 Flash Lite | 0.704 | 0.000235 | 2.56× |
| Llama 3.3 70B Turbo | 1.472 | 0.000491 | 5.35× |

A second pattern compounds the Chain-of-Thought paradox documented earlier in this section: across all three models, CoT prompting was not only no more accurate than Zero-Shot (and in DeepSeek's case 3.2 pp less accurate) but also between 12% and 25% more expensive per inference, due to the additional output tokens consumed by the reasoning trace. For tabular fact verification, then, Chain-of-Thought prompting is strictly dominated by Zero-Shot: it pays more to deliver no benefit or, in one model, a measurable loss.

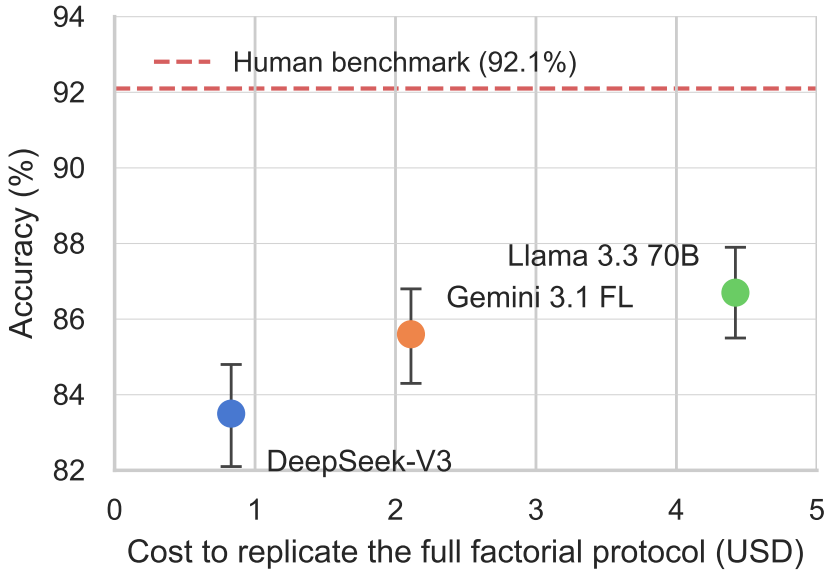


**Figure 3:** Cost-accuracy trade-off across the three evaluated LLMs on TabFact ($n = 8{,}986$ valid inferences). Error bars: Wilson 95% confidence intervals on accuracy. The dashed line marks the human benchmark (92.1%; [1]).

# 5. Discussion

## 5.1. The CoT Paradox in Structured Verification

Zero-Shot matched or outperformed CoT in all three models, a result that runs against the common assumption in the prompting literature [12, 13] that chain-of-thought reasoning improves performance across tasks, and that we read as a domain-specific effect. The tension is more apparent than real: those same demonstrations concentrated Chain-of-Thought's gains on arithmetic and multi-step reasoning, not on the direct evidence retrieval at issue here. Tabular fact verification is mostly a structured lookup, where the evidence sits in discrete cells and the answer rarely needs a multi-step argument, so asking the model to spell out intermediate steps adds room for error rather than catching it. The size of the penalty varies by model, with DeepSeek-V3 paying the steepest (3.2 pp) while Llama was unaffected (0.0 pp), which hints that the effect depends on the model: the more a model reasons out loud, the more chances it has to talk itself out of a correct cell lookup. In practice, adding CoT to an LLM-based verification pipeline can silently lower reliability instead of raising it.

## 5.2. Overconfidence as a Governance Risk

Among our findings, the calibration results carry the most direct implications for AI governance. The models display a pattern superficially reminiscent of the Dunning-Kruger signature [23] — expressed confidence decoupled from actual competence — though the analogy is behavioral, not cognitive: it arises here for mechanistically distinct, non-cognitive reasons, since the model has no self-model to misjudge. The evidence for this decoupling is consistent across metrics: $89.2\%$ of inferences were issued at maximum confidence regardless of correctness, yielding an ECE of $0.136$. On a 10-point scale, mean self-reported confidence differed by only $0.09$ points between correct ($9.90$) and incorrect ($9.81$) responses, and Gemini 3.1 Flash Lite showed no statistically significant discrimination between accurate and inaccurate predictions ($d = 0.17$, $p = 0.18$). This runs counter to the more optimistic view that language models largely know what they know [16]: in structured tabular verification, whatever self-assessment they exhibit elsewhere does not survive the demand for a discrete verdict. The operational consequence is the same either way: a verification component that reports maximum certainty across both correct and incorrect outputs gives a human reviewer no actionable signal for allocating scrutiny, the very function that well-calibrated uncertainty is meant to provide [17].

This calibration failure is separate from the accuracy gap, and it compounds it. A system with a known $13$–$17\%$ error rate is a quantifiable risk that can be priced into deployment, whereas a system that cannot reliably indicate which outputs fall in that margin turns a known risk into an operational blind spot. The asymmetry is most dangerous where it is least visible: in organizations lacking the personnel or processes for parallel verification, the model's miscalibrated confidence becomes, by default, the organization's confidence. Calibration is therefore not a secondary refinement to be addressed once accuracy is acceptable; in resource-constrained contexts it is the precondition for the model's accuracy to be safely usable at all.

## 5.3. Label Asymmetry and Audit Contexts

The ENTAILED/REFUTED asymmetry ($h = 0.22$) has underappreciated practical consequences. In audit or compliance contexts, the critical task is often *confirming* that a reported figure matches underlying data, not detecting fraud. In such settings, DeepSeek-V3's 75.8% accuracy on ENTAILED cases, despite 91.1% on REFUTED, would substantially underperform the aggregate headline figure of 83.5%. Practitioners selecting an LLM for a specific verification application should evaluate label-stratified accuracy, not only overall accuracy.

## 5.4. Cost-Effectiveness Trade-offs

The cost results reported in Section 4.9 reframe the accuracy comparison as a choice rather than a ranking. Llama 3.3 70B Turbo delivers the highest accuracy at the highest cost ($5.35\times$ DeepSeek-V3 for an additional 3.2 pp), Gemini 3.1 Flash Lite occupies an intermediate position on both axes, and DeepSeek-V3 defines the cost floor at the lowest accuracy. Which of these configurations is preferable cannot be answered in the abstract; it depends on the downstream cost of an incorrect verification relative to the marginal cost of each API call. For an organization where an undetected verification error triggers a regulatory response, a customer remediation cost, or a financial restatement, the accuracy premium of the more expensive model is easily justified. For an organization where the verification step is an internal convenience and downstream errors are caught by other controls, the cheapest model that meets the accuracy floor is the rational choice. The relevant analytical move, then, is to characterize the cost-of-error distribution of the use case and select the cheapest model whose expected error rate produces an acceptable expected loss, rather than to identify a single best model.

## 5.5. SME Governance Implications

Synthesizing the technical findings yields a coherent operational position for an SME deploying an LLM-based verification component. Three configuration decisions follow directly from the evidence. Default to Zero-Shot prompting for tabular verification tasks: in this study Chain-of-Thought was either neutral or actively harmful to accuracy across all three models, while always increasing per-inference cost; in the absence of an internal benchmark suggesting otherwise, the strictly dominated configuration should not be the default. Do not treat model-reported confidence as a quality control filter: with calibration error of the magnitude documented here, automated escalation rules keyed on model confidence will be operating on noise, and human reviewers should sample outputs independently of the confidence signal. Select model by cost-of-error rather than by headline accuracy: the $5.35\times$ cost difference between the cheapest and most expensive model studied here exceeds the differences in accuracy by an order of magnitude, so model selection should follow the use case's tolerance for incorrect verification rather than a generic benchmark ranking.

These decisions presuppose that the organization is in a position to generate its own evidence, however limited. The cost figures reported in Section 4.9 address that presupposition directly: at USD $2.45 for the full three-model experiment, and under USD $1.50 to evaluate any single model under the same design, replicating the protocol used here on a specific organizational corpus, rather than relying on TabFact, involves inference charges that any SME can absorb; the binding constraint lies elsewhere. In resource-constrained contexts, the barrier to evidence-based LLM governance is now less financial than analytic: it lies in the capacity to specify the verification task with enough precision that a small-scale factorial evaluation can produce decision-relevant numbers. That capacity is not trivial to build, but it is the appropriate target for institutional investment (in management training, in advisory services, in regulatory guidance) rather than infrastructure that larger organizations can afford and smaller ones cannot.

## 5.6. Limitations

Four limitations bound the scope of these findings. First, the DeepSeek API applies an internal temperature scaling ($T_{\text{model}} = T_{\text{api}} \times 0.3$), so temperature levels are not fully comparable across models. Second, the human reference (92.1%; [1]) was obtained under crowdsourced annotation conditions, and our stratification by linguistic/numeric type does not map directly onto TabFact's simple/complex channel split, which may affect direct comparison. Third, TabFact is English-only and Wikipedia-based, so generalization to domain-specific organizational tables requires further study. Fourth, 100 cases provide adequate statistical power for main effects but limit fine-grained subgroup analyses.

## 6. Conclusion

This study evaluated three commercially available LLMs (Gemini 3.1 Flash Lite, DeepSeek-V3, and Llama 3.3 70B Turbo) on the TabFact benchmark under a full factorial design, with the aim of producing deployment-relevant evidence calibrated to organizations without dedicated machine-learning operations capacity. All three models perform systematically below the human reference of $92.1\%$, with absolute accuracies between $83.5\%$ and $86.7\%$ across $8{,}986$ valid inferences.

Three technical findings inform the deployment decision. Chain-of-Thought prompting does not improve (and in one of the three models actively degrades) accuracy in tabular fact verification, contradicting an assumption commonly transferred from open-ended reasoning tasks; in addition, Chain-of-Thought increases per-inference cost by $12$–$25\%$ across models, leaving Zero-Shot strictly dominant in this task type. The three models exhibit severe miscalibration (ECE $= 0.136$), reporting maximum confidence in $89.2\%$ of inferences regardless of correctness, with the operational consequence that model-reported confidence cannot be used as a quality control filter. A systematic label-type asymmetry (Cohen's $h = 0.22$) means that aggregate accuracy may overstate reliability in applications where confirming true statements, rather than detecting false ones, is the primary verification objective.

A fourth finding bears on access to evidence-based governance. Reproducing the full factorial experiment ($9{,}000$ inferences) costs USD $2.45 in total, or USD $0.28–$1.47 per model, at current public list rates, and the cost ratios diverge from the accuracy ratios by roughly an order of magnitude: the most expensive model costs $5.35\times$ what the cheapest does for an additional $3.2$ pp of accuracy. Model selection, from an organizational standpoint, is therefore not a question of identifying the best model but of locating the cheapest model whose error rate is acceptable given the downstream cost of incorrect verification.

Three concrete extensions follow from the protocol established here. First, the factorial design is being applied to a second experimental layer of 12 deliberately ambiguous organizational decision scenarios (three operational domains $\times$ four difficulty levels, a planned 1,080-inference extension); because these scenarios lack ground-truth labels, the object of analysis shifts from correctness to the degree of divergence among the three models' decisions on the same input, a quantity with direct operational meaning for organizations whose decision quality depends on consistent automated judgment. Second, an annual longitudinal re-application of the protocol would turn this snapshot of a market in continuous evolution into a temporal series tracking how accuracy, calibration error, cost-effectiveness ratios, and cross-model agreement evolve in step with the commercial LLM industry itself. Third, both extensions admit expansion across the model and task dimensions: newer commercial and open-weight models, adjacent analytic operations such as reconciliation, anomaly explanation, and regulatory classification, and — of particular relevance to the Latin American context that frames this work — non-English corpora and domain-specific organizational tables.

Each of these directions extends the same methodological commitment that motivates this paper: producing evidence that organizations without dedicated machine-learning operations capacity can interpret, replicate, and adapt to their own decision contexts.

## Declaration on Generative AI

During the preparation of this work, the authors used Claude (Anthropic) for assistance in manuscript drafting, language polishing, structural review, and verification of statistical workflow code. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

## References

[1] W. Chen, H. Wang, J. Chen, Y. Zhang, H. Wang, S. Li, X. Zhou, W. Y. Wang, TabFact: A large-scale dataset for table-based fact verification, in: International Conference on Learning Representations

(ICLR), 2020, pp. 1–25. URL: https://openreview.net/forum?id=rkeJRhNYDH.

[2] World Economic Forum, Future of Jobs Report 2025, Technical Report, World Economic Forum, 2025. URL: https://reports.weforum.org/docs/WEF_Future_of_Jobs_Report_2025.pdf.

[3] D. Kreuzberger, N. Kühl, S. Hirschl, Machine learning operations (MLOps): Overview, definition, and architecture, IEEE Access 11 (2023) 31866–31879. doi:10.1109/ACCESS.2023.3262138.

[4] J. Schwaeke, A. Peters, D. K. Kanbach, S. Kraus, P. Jones, The new normal: The status quo of AI adoption in SMEs, Journal of Small Business Management 63 (2025) 1297–1331. doi:10.1080/00472778.2024.2379999.

[5] R. Durán, A. Moreno, S. Adasme, S. Rovira, V. Jordán, L. Poveda, Índice Latinoamericano de Inteligencia Artificial (ILIA) 2025, Documentos de Proyectos LC/TS.2025/68, Centro Nacional de Inteligencia Artificial (CENIA) and Comisión Económica para América Latina y el Caribe (CEPAL), Santiago, Chile, 2025. URL: https://repositorio.cepal.org/items/2646d237-7a4d-48f8-b7cf-e48ae2e5f9fe.

[6] J. Herzig, P. K. Nowak, T. Müller, F. Piccinno, J. M. Eisenschlos, TAPAS: Weakly supervised table parsing via pre-training, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4320–4333. doi:10.18653/v1/2020.acl-main.398.

[7] Q. Liu, B. Chen, J. Guo, M. Ziyadi, Z. Lin, W. Chen, J.-G. Lou, TAPEX: Table pre-training via learning a neural SQL executor, in: International Conference on Learning Representations (ICLR), 2022, pp. 1–19. URL: https://arxiv.org/abs/2107.07653.

[8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in Neural Information Processing Systems 33 (2020) 1877–1901.

[9] E. M. Bender, T. Gebru, A. McMillan-Major, M. Mitchell, On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), 2021, pp. 610–623. doi:10.1145/3442188.3445922.

[10] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, M. S. Bernstein, Generative agents: Interactive simulacra of human behavior, in: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23), 2023, pp. 1–22. doi:10.1145/3586183.3606763.

[11] S. Thapa, S. Shiwakoti, S. B. Shah, S. Adhikari, H. Veeramani, M. Nasim, U. Naseem, Large language models (LLM) in computational social science: Prospects, current state, and challenges, Social Network Analysis and Mining 15 (2025) 4. doi:10.1007/s13278-025-01428-9.

[12] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, Advances in Neural Information Processing Systems 35 (2022) 24824–24837.

[13] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, Advances in Neural Information Processing Systems 35 (2022) 22199–22213.

[14] X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, D. Zhou, Self-consistency improves chain of thought reasoning in language models, in: International Conference on Learning Representations (ICLR), 2023, pp. 1–24. URL: https://openreview.net/forum?id=1PL1NIMMrw.

[15] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, in: Proceedings of the 34th International Conference on Machine Learning (ICML), 2017, pp. 1321–1330.

[16] S. Kadavath, T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez, N. Schiefer, Z. Hatfield-Dodds, N. DasSarma, E. Tran-Johnson, et al., Language models (mostly) know what they know, arXiv preprint arXiv:2207.05221 (2022).

[17] U. Bhatt, J. Antorán, Y. Zhang, Q. V. Liao, P. Sattigeri, R. Fogliato, et al., Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty, in: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), 2021, pp. 401–413. doi:10.1145/3461702.3462571.

[18] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, et al., Ethical and social risks of harm from language

models, arXiv preprint arXiv:2112.04359 (2021).
[19] DeepSeek-AI, DeepSeek-V3-0324 model card, Hugging Face Model Hub, 2025. URL: https://huggingface.co/deepseek-ai/DeepSeek-V3-0324.
[20] J. L. Fleiss, Measuring nominal scale agreement among many raters, Psychological Bulletin 76 (1971) 378–382. doi:10.1037/h0031619.
[21] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, 2nd ed., Lawrence Erlbaum Associates, Hillsdale, NJ, 1988.
[22] E. B. Wilson, Probable inference, the law of succession, and statistical inference, Journal of the American Statistical Association 22 (1927) 209–212. doi:10.1080/01621459.1927.10502953.
[23] J. Kruger, D. Dunning, Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments, Journal of Personality and Social Psychology 77 (1999) 1121–1134. doi:10.1037/0022-3514.77.6.1121.