

# Pairwise Comparison Aggregation with Tool-Augmented LLMs for Procurement Value Lever Prioritization

Authors: Vitalii Shevchuk, Oleksandr Kondratiuk and Christoph Flöthmann

*akirolabs GmbH, Greifswalder Str. 208, 10405 Berlin, Germany*

Corresponding Author: [vitalii.shevchuk@akirolabs.com](mailto:vitalii.shevchuk@akirolabs.com)

Received: December, 2025    Published: March, 2026

## ARTICLE INFO

### Keywords:

Automated Decision Support;  
Comparison Aggregation; Large  
Language Models; Pairwise Comparison  
Aggregation; Procurement Value Lever  
Prioritization; Strategic Procurement

© 2026 by the Authors. This open-access article is distributed under a Creative Commons Attribution (CC-BY) 4.0 license, making research freely available to the public and supporting a greater global exchange of knowledge and human experiments.



## ABSTRACT

Traditional prioritization of procurement Value Levers (VLs) relies on manual expert evaluation across multiple strategic dimensions, making the process time-intensive, resource-demanding, and vulnerable to subjective inconsistency. This study investigates the use of Large Language Models (LLMs) to semi-automate VL ranking through a pairwise comparison aggregation framework enriched with established strategic tools. The proposed methodology integrates LLM-based contextual reasoning with the Kraljic Matrix, Procurement with Purpose, and SWOT analysis to produce consistent, context-aware rankings across a portfolio of 15 VLs. A systematic experimental design evaluates two model architectures (Llama-3.2-3B-Instruct and Llama-3.1-8B-Instruct), alternative prompt representations (numeric vs. categorical), varying contextual depth, and multiple aggregation mechanisms. Results show moderate alignment with proprietary baseline rankings (weighted Kendall's  $\tau = 0.38-0.45$ ), with performance improving through prompt calibration and strategic tool integration. Categorical representations consistently outperform numeric formats, indicating that qualitative descriptors better match LLM internal reasoning in procurement contexts. Strategic tool integration reduces ranking position variance by 84% and increases stability sixfold, while maintaining acceptable agreement with domain experts (68% exact agreement; Cohen's  $\kappa$  (quadratic) = 0.62). Position-specific variance is observed for a small subset of VLs (13–20%), which may reflect either model-induced structural effects or the capture of widely accepted procurement best practices. Overall, the findings demonstrate that LLM-based VL ranking is operationally viable for initial screening and sensitivity analysis, substantially reducing manual effort while preserving decision quality.

## 1.0 Introduction

Procurement operations have undergone significant transformation through digital automation and electronic solutions adoption. Traditional manual, paper based workflows are progressively replaced by software-enabled processes integrating data analytics and digital platforms. Empirical evidence demonstrates that electronic procurement (e-procurement) implementation enhances operational efficiency, improves supply chain performance, and facilitates sustainable procurement practices across diverse industrial sectors (Singh and Chan, 2022).

However, e-procurement adoption encounters substantial implementation barriers. Resistance to organizational change, technical infrastructure constraints, and supply-chain complexity before the deployment, particularly in sectors characterized by established traditional procurement frameworks (Gurgun et al., 2024).

Recent advances in Artificial Intelligence (AI), specifically LLMs and autonomous AI agents introduce novel capabilities for procurement transformation. AI-augmented procurement functions, including automated supplier inquiry, chatbots and algorithmic supplier selection

mechanisms, materially influence supplier behavior and pricing strategies through "automation plus intelligence" synergies (Cui, 2020).

Despite demonstrated potential, systematic reviews reveal limited LLM deployment in procurement contexts. Cross-industry analysis of LLM applications identifies procurement and supply-chain functions as substantially under-explored relative to internal operations or technology-centric organizational domains (Moenks et al., 2025). This gap indicates significant potential: LLMs remain underutilized in complex procurement tasks including dynamic supplier relationship management, risk-aware sourcing strategies, and multi-objective optimization.

This study addresses this research gap by investigating LLM application to procurement value lever prioritization through pairwise comparison aggregation frameworks. We propose an automated methodology integrating LLM contextual reasoning with established strategic procurement tools: KM, PwP, and SWOT Analysis to generate consistent, contextually grounded rankings across 15 VLs.

## 2.0 Literature Review

### 2.1 Digital Procurement and AI Integration

In a study on the evolution of electronic procurement, a four-stage evolutionary framework was proposed characterizing procurement transformation: Manual Operations (MO) → Electronic Procurement (EP) → Sustainable Procurement (SP) → Intelligent Procurement (IP). The intelligent procurement paradigm integrates sustainability principles, process automation, AI-augmented decision support, advanced analytics, and smart contract mechanisms-establishing a foundation for procurement systems characterized by enhanced efficiency, transparency, sustainability compliance, and adaptive capability in response to dynamic operational requirements (Chan and Owusu, 2022). This demonstrates significant emphasis on AI and integration of recent technologies in IP.

Despite this progress, numerous studies highlight that adoption of IP remains uneven and slow: despite increasing scholarly interest, many firms and public agencies still operate largely manually due to conservative culture, steep learning curves, cost concerns, and organizational resistance (Chan and Owusu, 2022; Afolabi et al., 2019).

Several fundamental limitations prevent rapid LLM adoption in complex procurement decision-making contexts. Hallucination phenomena, where models generate plausible yet factually incorrect outputs, could yield significant risk in domains where errors in prioritization lead to material and financial consequences (Ye et al., 2025). General-purpose LLMs lack grounding in domain-specific frameworks, potentially producing recommendations misaligned with established procurement methodologies (Moenks et al., 2025). The opaque nature of LLM reasoning conflicts with procurement auditability and compliance requirements, where decision justification is frequently mandatory (Pesch et al., 2025). Calibration studies reveal LLM overconfidence patterns: models express high certainty even for uncertain predictions, a problematic characteristic for risk-aware sourcing strategies. Output inconsistency across identical or almost identical prompts undermines the reproducibility essential for systematic procurement processes.

Additional challenges compound these limitations. Numerical reasoning weaknesses constrain precise quantitative trade-off evaluation. The tacit knowledge problem, where expert procurement judgment involves implicit knowledge that is difficult to encode in prompts, further limits direct LLM application to strategic prioritization tasks.

These limitations collectively explain why, despite demonstrated potential in adjacent domains, LLM deployment in strategic procurement requiring expert judgment remains substantially limited. Regardless, AI introduces major opportunities across procurement-from digitizing and automating tender- and bidding-related processes (making procedures faster and less error-prone) to enabling data-driven decision support, risk detection, and

improved supplier evaluation - all of which can boost efficiency, transparency and accountability in public procurement (Aboelazm and Dganni, 2025). By leveraging AI's ability to analyze large datasets (past bids, supplier performance, market conditions), procurement bodies can streamline supplier selection, forecast demand, spot fraud or irregularities early, ensure compliance, optimize spend and ultimately convert procurement from a purely transactional process into a strategic, intelligent function (Andhov et al., 2025).

## 2.2 LLMs for Ranking and Decision Tasks

LLMs demonstrate increasing deployment in ranking and judgment tasks, enabling scalable evaluations where human assessment proves resource-intensive (Anghel et al., 2025). Pairwise meta-evaluation frameworks indicate LLMs function as reliable comparative judges when outputs undergo systematic aggregation, while revealing bias patterns requiring mitigation in decision pipelines (Anghel et al., 2025). Multi-judge prompting, rubric-guided evaluation, and tool-augmented reasoning substantially improve consistency and alignment with expert judgments (Gu et al., 2025).

However, LLM reasoning capabilities exhibit domain-specific performance disparities. Comparative evaluations report reliable performance on verbal and categorical reasoning, yet degraded accuracy on numeric and arithmetic operations (Abdelkarim et al., 2025). Models achieve better scores on verbal items relative to numerical tasks, indicating comparative strengths in qualitative, ordinal, or descriptive criterion processing rather than precise quantitative computation. For structured decision contexts, these findings establish that LLM based systems achieve optimal reliability when evaluation criteria use categorical representations, while decisions requiring exact numerical aggregation necessitate hybrid human-AI workflows.

## 2.3 Evaluation Metrics for Ranking Systems

Ranking system evaluation in procurement VL prioritization contexts draws methodological foundations from established practices in Information Retrieval (IR) and recommender systems research. The literature distinguishes between primary performance metrics assessing ranking quality and secondary stability metrics quantifying output consistency under input perturbations.

**2.3.1 Primary Ranking Metrics.** IR research uses several standardized metrics for ranked output assessment, particularly when relevance exhibits gradation or ordering accuracy constitutes the primary evaluation criterion. Widely adopted metrics include Normalized Discounted Cumulative Gain (NDCG), Mean Average Precision (MAP).

NDCG accommodates graded relevance structures while applying position-dependent penalties: items appearing lower in rankings are subject to greater discounting. This characteristic proves particularly valuable for top-heavy applications where early-position accuracy disproportionately influences system utility. Comparative reviews demonstrate that NDCG is a better metric to pick over binary relevance metrics in contexts requiring nuanced relevance discrimination (Baeza-Yates and Ribeiro-Neto, 2011).

MAP aggregates precision across varying recall thresholds, averaged across query sets or decision tasks. This metric emphasizes consistent relevant item retrieval rather than isolated query success. IR literature identifies MAP as the predominant metric for cross-query system performance evaluation (Ricci et al., 2011).

These metrics supersede simpler binary relevance measures including precision and recall, when used independently fail to capture the significance of ordering within ranked results.

**2.3.2 Secondary and Stability Metrics.** Primary metrics quantify overall effectiveness yet may obscure ranking system behavior under input perturbations: scenarios where minor input modifications induce substantial rank fluctuations. Stability-oriented and consistency metrics address this evaluation gap.

Rank correlation measures, particularly Kendall's tau (Yilmaz, 2007), enable ranking stability assessment across temporal periods or data subsets. This analytical framework facilitates evaluation of ranking method robustness against temporal dynamics or data variability - considerations particularly relevant in domains characterized by evolving input parameters including procurement criteria, vendor data, and market conditions.

Complementary metrics quantifying position displacement: measuring prediction shift magnitude relative to ground truth or rank variance across experimental runs (Webber, 2010) enable consistency and reproducibility assessment. While classical IR literature lacks explicit standardization of position displacement or intra-experiment consistency terminology, correlation based and stability-aware evaluation methodologies effectively serve analogous functions: revealing ranking output divergence even when aggregate metrics (MAP, NDCG) remain comparable.

**2.3.3 Metric Selection Rationale for Procurement Applications.** This study focuses on Kendall's tau rather than conventional IR metrics based on methodological alignment considerations. The preference arises from task-specific evaluation objectives: procurement VL prioritization requires ordering quality assessment rather than graded relevance retrieval performance measurement.

Metric selection depends critically on which ranking aspects prove most meaningful for the specific application domain.

For procurement contexts where the primary objective involves comparing relative orderings generated by distinct models or aggregation mechanisms Kendall's tau demonstrates structural alignment with problem requirements. This non-parametric measure directly quantifies ordinal association through concordant versus discordant pair evaluation, providing interpretable assessment of ranking correspondence without imposing assumptions regarding relevance score distributions or position-dependent utility functions characteristic of NDCG or MAP frameworks.

## 3.0 Methodology

### 3.1 General experiment design flow

Research Objective Statement: the primary objective of this study is to develop an automated methodology for ranking procurement VLs through LLMs integration with qualitative business objective profiles. The proposed framework aims to incorporate contextual information including value-level specifications and supporting tool data for cross-validation purposes.

The methodology must demonstrate computational efficiency and parameter sensitivity while maintaining output consistency comparable to expert-validated procurement solutions. This approach addresses the challenge of systematic VL prioritization in procurement operations, where traditional manual ranking methods prove time-intensive and susceptible to subjective bias.

By leveraging LLMs capability for contextual understanding and multi-dimensional analysis, the framework seeks to automate decision support processes while preserving the accuracy standards established through domain expert validation.

General experiment flow goes throughout stages:

- 1) Preparing input configuration;
- 2) Generating possible non-repeating combinations of VLs pairs;
- 3) Forming prompt from templates using input configuration;
- 4) Predicting all pairs with LLMs;
- 5) Aggregating LLM votes for each VL;
- 6) Sorting final VL ranks with respect to votes.

### 3.2 Data Collection and Processing

Input data for this research consists of:

1) Variations of scenario information that define the desired output of VL rank that is provided in Table 1. Where Focus could be from 1 to 10 and Category represents the focus for procurement strategy. Also, only minimum of 3 and maximum 7 Categories could be selected at once;

2) Output ranks of VL (only 15 related to PwP) for a specific input parameters and tools. This data is derived from akirolabs API and provided in Table 2 as output example;

3) 15 PwP focused VLs. List of VL related to PwP with metadata and focus dimensions values defined by procurement experts. Examples are provided in Table 5.

### 3.3 Baseline System Validation

A critical methodological concern involves establishing the validity of the v1 proprietary baseline used as ground truth throughout this study. Unlike LLM experiments where ranking decisions emerge from opaque model inference, the v1 baseline represents a deterministic algorithmic implementation informed by domain expertise. This section documents the validation evidence supporting v1 as a legitimate benchmark.

Table 1. Example of input scenario configuration

Category	Focus	Selected
Savings Margin	1	✗
Sustainability	1	✗
Resilience	5	✓
Agility	7	✓
Innovation	1	✗
Quality	1	✗
Growth	1	✗
Regulatory	8	✓
Diversity	1	✗
Efficiency	1	✗
Time to Market	1	✗

1) Expert-Informed Development: The v1 ranking algorithm was developed in collaboration with five procurement domain experts, each possessing over 10 years of experience in strategic procurement and supply chain management. The algorithm design, VL metadata weights, and scenario-response mappings underwent iterative refinement based on expert feedback across multiple development cycles. This expert involvement ensures that v1 encodes established procurement domain knowledge rather than arbitrary computational rules;

2) Commercial Deployment Validation: The v1 system has been deployed in commercial procurement operations for over two years, serving more than 10 enterprise clients across diverse industry sectors. This extended operational history provides implicit validation: systematic ranking failures would have surfaced through client feedback and operational metrics during commercial use. While commercial deployment does not substitute for controlled validation studies, it establishes baseline viability for real-world procurement decision support;

3) Independent Expert Cross-Validation Study: To address concerns regarding ground truth validity, we conducted a cross-validation study with two independent procurement domain experts (each with 10+ years of experience, neither involved in v1 development). Due to resource constraints, direct validation of exact 15-position rankings across all 965 scenarios was infeasible such validation would require experts to assess 14,475 individual VL positions. Instead, we used a coarser-grained evaluation protocol using a 5-point holistic ranking quality scale as shown in Table 3.

Table 2. Example of VL rank output after processing

Rank	Procurement Initiative
1	Local sourcing
2	Virtual supplier meetings
3	Whistleblowing application
4	Re-use and recycling
5	Net-zero carbon emission supply chain
6	Circular procurement
7	Sustainability targets in contracts
8	Durable products
9	Diverse / minority-owned supplier support
10	Health and safety
11	Sustainable supplier awards
12	Resource consumption reduction
13	Sustainability compliance
14	Waste reduction
15	Product and service sharing and re-use

Table 3. Ranking Quality Assessment Scale

Score	Description
1	Mostly wrong; top ranks incorrect
2	Some top items correct; many misplacements
3	About half correct; moderate errors
4	Most items are correct; minor swaps
5	Perfect or nearly perfect ranking; all top items correct

A stratified random sample of 100 scenarios (10.4% of the full dataset) was selected to ensure representation across input parameter combinations. Each scenario's v1-generated ranking was independently assessed by both experts using the 5-point scale.

Inter-Annotator Agreement Analysis: We used Cohen's Kappa ( $k$ ) to quantify inter-annotator reliability, as this metric accounts for agreement occurring by chance unlike simple percent agreement, which can be misleadingly high when rating categories are imbalanced;

Cohen's Kappa is defined as:

$$k = \frac{P_o - P_e}{1 - P_e} \quad (1)$$

where  $P_o$  represents observed agreement (proportion of cases where annotators agree) and  $P_e$  represents expected agreement by chance (based on marginal distributions of each annotator's ratings).

For ordinal scales such as our 5-point ranking quality assessment, weighted kappa variants are preferred because they penalize disagreements proportionally to their magnitude. Linear weighted  $\kappa$  treats disagreements linearly (off-by-2 is twice as problematic as off-by-1), while quadratic weighted  $\kappa$  penalizes larger disagreements more heavily (off-by-2 is four times as problematic as off-by-1). Quadratic weighting aligns with the intuition that adjacent ratings (e.g., 4 vs 5: "most items correct" vs "perfect") share meaningful similarity, whereas distant ratings (e.g., 1 vs 5: "mostly wrong" vs "perfect") represent fundamental assessment disagreement.

Following the widely adopted interpretation scale (Landis & Koch):

- $k < 0.00$ : Poor;
- $k = 0.00\text{--}0.20$ : Slight;
- $k = 0.21\text{--}0.40$ : Fair;
- $k = 0.41\text{--}0.60$ : Moderate;
- $k = 0.61\text{--}0.80$ : Substantial;
- $k = 0.81\text{--}1.00$ : Almost Perfect.

Table 4. Inter-Annotator Agreement Metrics for Baseline Validation

Metric	Value	Interpretation
Total Samples	100	—
Exact Agreement Rate	68.0%	—
Cohen's $k$ (unweighted)	0.4422	Moderate
Cohen's $k$ (linear weighted)	0.5164	Moderate
Cohen's $k$ (quadratic weighted)	0.6166	Substantial
Mean Absolute Difference	0.330	—
Off by $\leq 1$ Rate	99.0%	Nearly all within $\pm 1$

Table 5. Example of 1 out of 15 VL metadata

Field	Value
VL ID	12
VL Type	Commercial
VL	Local sourcing
Demand bundling	-
Need reduction	-
Technical specification	-

Field	Value
Competition	1
Supply base management	1
Supplier relationship management	-
Supply chain operations	-
PwP	1
Savings / Margin	2
Sustainability	8
Innovation	2
Agility	7
Resilience / Securing supply	7
Quality	5
Growth	2
Regulatory	5
Diversity	2
Efficiency	2
Time to market	5
Hint	Best applicable if total costs are driven by transportation, and specialized suppliers are clustered around your own facilities
Ease of Implementation	Medium
Time to Impact	Short-term
Cross-Check Tools	SWOT, PwP, etc.

The quadratic weighted kappa ( $\kappa = 0.62$ ) achieves "substantial" agreement, with 99% of disagreements falling within  $\pm 1$  rating point. This indicates that while experts occasionally differed on whether a ranking was "perfect" versus "mostly correct," they rarely disagreed on whether rankings were fundamentally acceptable. The unweighted  $\kappa = 0.44$  ("moderate") reflects stricter exact-match criteria, while the progression from unweighted to quadratic-weighted  $\kappa$  ( $0.44 \rightarrow 0.52 \rightarrow 0.62$ ) confirms that disagreements predominantly occur between adjacent categories rather than distant ones as can be seen withing Table 4.

5) Validation Limitations: Several limitations constrain the strength of these validation claims. First, the 5-point scale represents coarse-grained assessment that cannot detect subtle position-specific errors. Second, the 100-sample evaluation covers only 10.4% of experimental scenarios. Third, two expert assessors, while experienced, represent a limited validation panel. Fourth, the experts assessed v1 outputs in isolation rather than comparative evaluation against LLM alternatives. Despite these constraints, the validation study provides empirical evidence that v1 rankings achieve acceptable quality by domain expert standards, establishing v1 as a defensible, though imperfect, benchmark for comparative evaluation of LLM based alternatives.

This multi-layered validation approach, expert-informed development, extended commercial deployment, and independent cross-validation with substantial inter-annotator agreement

addresses the “black box comparison” concern by demonstrating that the v1 baseline, while proprietary, produces rankings that domain experts judge as acceptable across sampled scenarios as shown in Fig 1.

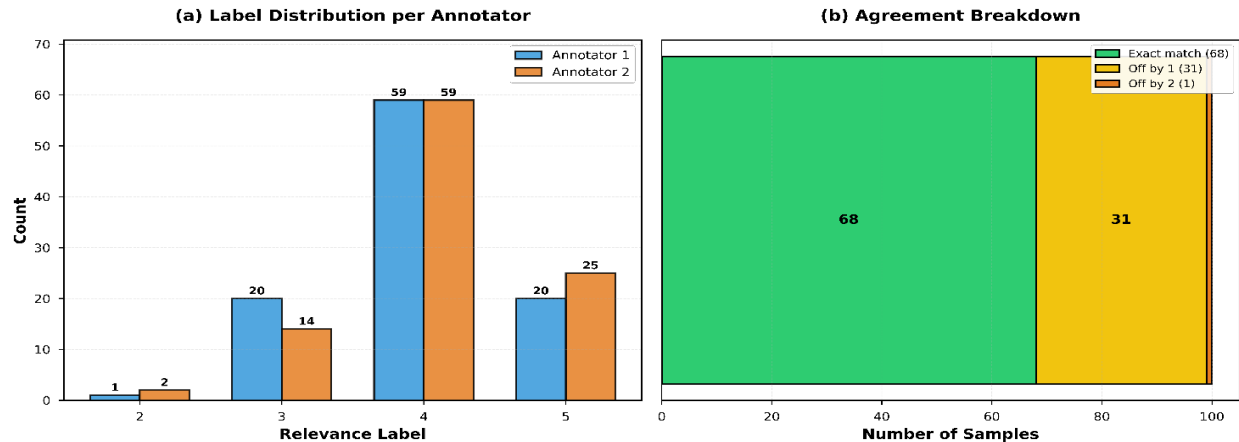


Fig 1. Label distribution per annotator (left) agreement breakdown (right)

### 3.4 Strategic Tool Integration

The experimental framework incorporates established procurement strategy tools to provide structured contextual information for VL ranking. Three strategic instruments were selected based on their direct relevance to the PwP framework and their capacity to influence VL prioritization decisions.

Tool Selection and Configuration:

- **Kraljic Matrix:** A portfolio segmentation tool that classifies procurement items across two dimensions: supply risk and profit impact to guide differentiated sourcing strategies. This matrix enables strategic resource allocation by identifying items requiring enhanced supplier relationship management or risk mitigation initiatives;
- **Procurement with Purpose Framework:** A strategic methodology that systematically integrates sustainability criteria, social value considerations, and ethical procurement principles into purchasing decisions. This framework ensures alignment between procurement activities and organizational objectives related to environmental goals and corporate social responsibility;
- **SWOT Analysis:** A diagnostic instrument used to evaluate internal organizational capabilities (strengths and weaknesses) in conjunction with external market dynamics (opportunities and threats) within the procurement context. This analysis facilitates strategy formulation by identifying strategic fit between internal competencies and external conditions;

**Expert Validation and Tool Configuration:** Each strategic tool underwent manual configuration by procurement domain specialists to ensure accuracy and relevance to the experimental context. Tool parameters remained fixed throughout the experiments unless explicitly modified for comparative analysis purposes. While comprehensive procurement strategies typically incorporate up to 12 analytical tools, the selected three- tool configuration represents the instruments most directly influencing PwP-related VLs, thereby ensuring experimental focus and analytical tractability.

**Tool-Category Alignment Rationale:** The selection of KM, PwP Framework, and SWOT Analysis was not arbitrary but driven by their direct methodological connection to PwP-related VLs. These three tools fully cover the analytical dimensions relevant to sustainability-focused procurement decisions without requiring additional frameworks. Alternative VL categories would necessitate incorporating tools such as Risk Management,

PESTLE analysis, Cost Driver decomposition, and supplier-specific assessment – instruments substantially increasing experimental complexity while introducing dependencies on supplier data that falls outside this study's scope.

- **Data Sampling Strategy:** The complete input parameter space comprises several million possible combinations when considering all tool configurations and VL interactions. To maintain computational feasibility while ensuring statistical validity, a stratified sampling approach was used, generating 965 representative data points for analysis;
- **VL Category Selection Rationale:** Operational procurement environments organize VLs into distinct strategic categories, with category composition varying across organizations. The following categories were evaluated for experimental suitability (Table 6).

Table 6. VL Category Evaluation for Experimental Selection

Category	VL Count	Selection Decision	Rationale
Supply Chain Operations	30+	Excluded	Excessive scope for paper; would require 435+ pairwise comparisons per sample, introducing prohibitive computational overhead
Procurement with Purpose (PwP)	15	Selected	Optimal balance: sufficient complexity for methodology validation (105 pairwise comparisons) while maintaining computational tractability
Supplier Relationship Management	16	Excluded	Comparable VL count but requires supplier-specific data integration and dedicated supplier assessment tools not within study scope
Technical Specifications	16	Excluded	Requires domain-specific technical expertise for tool configuration; less generalizable findings
Competition	7	Excluded	Insufficient VL count for meaningful ranking analysis; limited statistical power for detecting variance patterns
Demand Bundling	6	Excluded	Insufficient VL count; pairwise comparison framework loses discriminative power with <10 items
Need Reduction	4	Excluded	Minimal VL count precludes robust ranking methodology evaluation

PwP was selected based on quantitative factors (optimal VL count for computational feasibility and statistical validity) combined with tool ecosystem alignment. The KM, PwP Framework, and SWOT Analysis instruments fully cover PwP-related VLs without requiring additional analytical tools or external data dependencies. Categories such as Supplier Relationship Management and Technical Specifications, while offering comparable VL counts, would necessitate incorporating Risk Management frameworks, PESTLE analysis, Cost Driver decomposition, and supplier performance data—substantially complicating experimental design while introducing confounding variables that would obscure the primary research questions regarding LLM ranking behavior.

This experimental configuration balances practical applicability with research tractability, enabling systematic evaluation of the automated ranking methodology under realistic yet manageable conditions. The methodology and findings are expected to generalize to other VL categories, though category-specific tool integration and potential supplier data requirements would need to be addressed in future work.

### 3.5 Large Language Model Configuration

Initial evaluation considered proprietary LLMs, but inference costs proved prohibitive even for limited experimental runs. This constraint directed focus toward open-source alternatives. Among available options, the Llama model family emerged as the optimal

choice based on capability and computational efficiency. Full list of factors and models can be found in the Table 7.

Table 7: Multi-Factor models assessment

Model	Size (parameters / typical)	Cost category for input /output tokens	Latency / performance estimate	Quality (reasoning capability / use-case fit)
GPT-4.5	~780 billion parameters (estimated)	Very high	Low (fast inference)	Very high (advanced reasoning; suitable for complex procurement analytics, scoring)
Claude	Size undisclosed publicly; some estimates: Claude 3.5 Sonnet ">175 billion" parameters	Moderate	Low (fast inference)	High (strong summarization, compliance review, communication-heavy procurement tasks)
Llama- 3.2-1B-instruct	~ 1 billion parameters	Ultra low	Ultra low	Bad-to-moderate (suitable for simple automation, routing, basic classification)
Llama- 3.2-3B-instruct	~ 3 billion parameters	Low	Very low	Moderate (suitable for lightweight assistants, internal procurement workflows)
Llama- 3.1-8B-instruct	~ 8 billion parameters	Low-to-moderate	Low	Good (capable of mid-complexity tasks like summaries, drafting, supplier review)
Llama- 3.1-14B	~ 14 billion parameters	Moderate	Medium	Good-to-High (strong open-source model for analysis, documentation, internal procurement reasoning)

Model size selection involved systematic evaluation across the Llama parameter range:

- 1B models: Excluded due to insufficient capacity for complex multi-step reasoning required in VL ranking tasks;
- 3B models: Demonstrated adequate performance but represented the lower threshold of acceptable quality for procurement-specific analysis;
- 8B models: Identified as optimal, providing strong performance while maintaining practical inference speeds and manageable resource requirements;
- 14B+ models: Offered marginal quality improvements but introduced excessive computational overhead.
- Based on this analysis, two models were selected for experiments:
- Llama-3.2-3B-instruct: Baseline model for performance comparison;
- Llama-3.1-8B-instruct: Primary model based on optimal quality-to-efficiency ratio.

This configuration enabled direct comparison across model capacities while maintaining computational feasibility.

## 4.0 Experimental Setup

### 4.1 Prompt Evolution

In order to explore and investigate effect of prompts and various data inputs and how they affect the output ranking we designed a series of experiments.

## 1) Original proprietary VL implementation (Experiment v1);

As input this algorithm expects variations of scenario information. Under the hood it operates with numeric values for a fixed VL table.

## 2) LLM based solution Llama-3.2-3B-instruct numeric values with VL metadata. (Experiment v2.1);

This solution has 2 main points:

- 1 - is LLM that is tasked of choosing what VL is better in current situation;

Prompt Example:

You are given:

- A **Task** description.
- 2 items of **Value Lever** with short descriptions and additional information.
- A business objective profile (based on procurement experts' observations) for each of the value levers on a 1–9 scale:  
"1 - No relevance", "2 - Very low priority", "3 - Low priority", "4 - Emerging focus", "5 - Moderate priority", "6 - Growing strategic relevance", "7 - High importance", "8 - Very high importance", "9 - Critical to strategy".
- A **Scenario information** section describing the context in which the value levers are to be evaluated.
- A section on **Strategy tools information** that can be used to guide your assessment.

---

Your job is to:

- Evaluate how relevant each Value Lever is to the task and scenario taking into account business objective profile.
- Decide whether the **first Value Lever** is **more important** than the **second one** with respect to the task and scenario.
- Only output one word: **Yes** if the first is more important, **No** if the second is more important or they are equal.

---

### Scenario information:

This list outlines the primary focus areas for a given procurement strategy scenario.

- Selected: True — This key area is a main focus in the current procurement strategy scenario.
- Selected: False — This key area is not important and excluded from the current scenario.

Focus Level:

Each key area is assigned a focus level on a scale from 1 to 5, where:

1 = Lowest priority

5 = Highest priority

Actual values:

key: tool.valueLeversStrategic.question.savingsMargin

selected: True

focus: 4

key: tool.valueLeversStrategic.question.sustainability

selected: True

focus: 5

key: tool.valueLeversStrategic.question.resilience

selected: False

focus: 1

key: tool.valueLeversStrategic.question.agility

selected: True

focus: 4

key: tool.valueLeversStrategic.question.innovation

selected: True

focus: 2

key: tool.valueLeversStrategic.question.quality

selected: True

focus: 2

key: tool.valueLeversStrategic.question.growth

selected: False

focus: 1

key: tool.valueLeversStrategic.question.regulatory

selected: True

focus: 5

key: tool.valueLeversStrategic.question.diversity

selected: False

focus: 1

key: tool.valueLeversStrategic.question.efficiency

selected: True

focus: 5

key: tool.valueLeversStrategic.question.timeToMarket

selected: False

focus: 1

### First Value Lever with business objective profile:

'Value Lever Name': 'Local sourcing'

'Addressable Business Objective': ['Resilience', 'Agility', 'Sustainability']

'Hint': 'Best applicable if total costs are driven by transportation and specialized suppliers are clustered around your own facilities'

'Savings/Margin': 2

'Sustainability': 8

'Innovation': 2

'Agility': 7

'Resilience/Securing supply': 7

'Quality': 5

'Growth': 2

'Regulatory': 5

'Diversity': 2

'Efficiency': 2

'Time to market': 5

### Second Value Lever with business objective profile:

'Value Lever Name': 'Virtual supplier meetings'

'Addressable Business Objective': ['Sustainability', 'Efficiency']

'Hint': 'Reducing carbon emissions by traveling less adds up over time'

'Savings/Margin': 2

'Sustainability': 7

'Innovation': 2

'Agility': 2

'Resilience/Securing supply': 2

'Quality': 2

'Growth': 2

'Regulatory': 2

'Diversity': 2

'Efficiency': 7

'Time to market': 2

---

### Output format:

Yes

(or)

No

Do not add any explanation or extra text. Just return `Yes` or `No`.

...

2 - is a combination of all possible pairs of 15 VL in a way that they do not repeat. Summary of best rank is picked using most votes.

3) LLM based solution Llama-3.2-3B-instruct numeric values with VL metadata and scenario values calibration. (Experiment v2.2);

Changed part of the prompt example:

...

You are given:

- A **Task** description.
- 2 items of **Value Lever** with short descriptions and additional information.
- A business objective profile (based on procurement experts' observations) for each of the value levers using qualitative priority levels:  
"Very Low" (1-2), "Low" (3-4), "Medium" (5-6), "High" (7-8), "Critical" (9).
- A **Scenario information** section describing the context in which the value levers are to be evaluated.

Each Value Lever includes a business objective profile, where numeric scores (1–9) are converted into qualitative priority levels:

1-2 = Very Low

3-4 = Low

5-6 = Medium

7-8 = High

9 = Critical

Key Area	Selected	Focus Level
Savings/Margin	True	Low
Sustainability	True	Medium
Resilience	False	Minimal
Agility	True	Low
Innovation	True	Minimal
Quality	True	Minimal
Growth	False	Minimal
Regulatory	True	Medium
Diversity	False	Minimal
Efficiency	True	Medium
Time to market	False	Minimal

### First Value Lever with business objective profile:

'Value Lever Name': 'Local sourcing'

'Addressable Business Objective': ['Resilience', 'Agility', 'Sustainability']

'Hint': 'Best applicable if total costs are driven by transportation and specialized suppliers are clustered around your own facilities'

'Savings/Margin': Very Low

'Sustainability': High

'Innovation': Very Low

'Agility': High

'Resilience/Securing supply': High

'Quality': Medium  
 'Growth': Very Low  
 'Regulatory': Medium  
 'Diversity': Very Low  
 'Efficiency': Very Low  
 'Time to market': Medium  
 etc...  
 ...

Key difference is that we replace all numeric values with category based information. The logic is that LLM would perform with better latent representations in this case. Other experimental conditions were held constant.

4) LLM based solution Llama-3.2-3B-instruct numeric values with VL metadata and scenario values calibration and PwP tools information. (Experiment v2.3);

Changed part of the prompt example:

...

### Strategy tools information:

PwP:

Factor: E-CO2 footprint & emissions, Dimension: Environment, Title: Energy Consumption, Description: Datacenter consumption increased by 50%, Company: , Impact: 30

Factor: E-Renewable energy & power efficiency, Dimension: Environment, Title: Energy Typ (Renewable =, Description: Green DC / renewable energy - overcompensating the high consumption, Company: , Impact: 90

Factor: S-Human rights & employee well-being, Dimension: Social, Title: Local Human Rights , Description: have to accept local circumstances , Company: , Impact: 40

Factor: S-Society & public matters, Dimension: Social, Title: Political change ,De-scription: contributing to political change , Company: , Impact: 70

Factor: S-Health & safety, Dimension: Social, Title: Higher standards, Description: Certified workplace / provider, , Company: , Impact: 80

E-Pollution, waste & recycling, Dimension: Environment, Title: Environmental SDG 2030 , Description: Agenda for 2030 goals in the selected dimension, Company: EcoLead, Impact: 60

SWOT: \*\*SWOT Analysis\*\*

### GATHER DATA (Input Elements)

\*\*Internal Factors - Strengths:\*\*

\* Supplier network

→ vs Quantum Computing (opportunity): 4/4

→ vs SaaS (opportunity): 4/4

→ vs New competitors (threat): 0/4

→ vs GDPR (threat): 4/4

\* SCM expertise

→ vs Quantum Computing (opportunity): 0/4

- vs SaaS (opportunity): 0/4
- vs New competitors (threat): 2/4
- vs GDPR (threat): 0/4
- \* New corporate strategy
  - vs Quantum Computing (opportunity): 3/4
  - vs SaaS (opportunity): 2/4
  - vs New competitors (threat): 0/4
  - vs GDPR (threat): 0/4
- \*\*Internal Factors - Weaknesses:\*\*
  - \* Long product lifecycle
    - vs Quantum Computing (opportunity): 1/4
    - vs SaaS (opportunity): 1/4
    - vs New competitors (threat): 1/4
    - vs GDPR (threat): 0/4
  - \* Fragmented processes
    - vs Quantum Computing (opportunity): 0/4
    - vs SaaS (opportunity): 4/4
    - vs New competitors (threat): 2/4
    - vs GDPR (threat): 1/4
  - \* Employee turnover
    - vs Quantum Computing (opportunity): 0/4
    - vs SaaS (opportunity): 0/4
    - vs New competitors (threat): 0/4
    - vs GDPR (threat): 0/4
- \*\*External Factors - Opportunities:\*\*
  - \* Quantum Computing
  - \* SaaS
- \*\*External Factors - Threats:\*\*
  - \* New competitors
  - \* GDPR
- ### INSIGHTS (Scored Analysis)
  - \*\*STRENGTHS (Internal Positive Factors):\*\*
    - \* Supplier network: 12/16 (75.0%)
    - \* SCM expertise: 6/16 (37.5%)
    - \* New corporate strategy: 5/16 (31.2%)
  - \*\*WEAKNESSES (Internal Negative Factors):\*\*
    - \* Long product lifecycle: 3/16 (18.8%)

\* Fragmented processes: 7/16 (43.8%)

\* Employee turnover: 0/16 (0.0%)

\*\*OPPORTUNITIES (External Positive Factors):\*\*

\* Quantum Computing: 8/24 (33.3%)

\* SaaS: 15/24 (62.5%)

\*\*THREATS (External Negative Factors):\*\*

\* New competitors: 7/24 (29.2%)

\* GDPR: 8/24 (33.3%)

### STRATEGIC SUMMARY

\*\*Internal Analysis:\*\*

\* Total Strengths: 23/48 (47.9%)

\* Total Weaknesses: 10/48 (20.8%)

\* Internal Balance: Strength-focused

\*\*External Analysis:\*\*

\* Total Opportunities: 23/48 (47.9%)

\* Total Threats: 15/48 (31.2%)

\* External Balance: Opportunity-rich

\*\*Strategic Positioning:\*\*

\* Overall Position: Aggressive Strategy (Leverage strengths to capitalize on opportunities)

\*\*Key Action Guidance:\*\*

\* Leverage Supplier network strength

---

''

The experimental prompt incorporates RAG-inspired logic to enhance model context with relevant strategic tool information. This approach ensures that VL ranking decisions are grounded in applicable procurement frameworks while avoiding irrelevant information injection.

#### Tool Content and Configuration Details:

- Kraljic Matrix: Configured with organization-specific supply risk assessments (1-4 scale) and profit impact ratings for procurement categories. The matrix classifies items into four quadrants (Strategic, Leverage, Bottleneck, Non-critical), with each VL mapped to relevant quadrant characteristics influencing prioritization decisions;
- PwP Framework: Contains sustainability factor assessments across Environment (E), Social (S), and Governance (G) dimensions. Each factor includes: dimension classification, descriptive title, impact assessment (0-100 scale), and organizational context. Example factors shown in prompts include CO2 footprint, renewable energy usage, human rights considerations, and health/safety standards;
- SWOT Analysis: Structured as internal factors (Strengths, Weaknesses) cross-referenced against external factors (Opportunities, Threats) with pairwise relevance scores (0-4 scale). The tool generates aggregate scores for each factor

category and derives strategic positioning recommendations (Aggressive, Defensive, Conservative, Competitive) that inform VL prioritization context;

- Context Selection Logic: Tool information is conditionally included in model prompts based on the intersection of tools relevant to both VLs being compared. This intersection-based approach is critical for maintaining unbiased pairwise comparisons;
- Intersection Based Tool Injection: For each pairwise VL comparison, the system identifies which strategic tools are referenced by each VL and computes their intersection. Only tools appearing in both VLs' reference sets are included in the prompt context;
- Example: Consider a comparison between VL\_A (references: Kraljic Matrix, PwP Framework) and VL\_B (references: Kraljic Matrix, PwP Framework, SWOT Analysis). The injected context includes only Kraljic Matrix and PwP Framework information- SWOT Analysis is excluded despite being relevant to VL\_B;
- Rationale for Intersection Approach: Including tools referenced by only one VL would introduce asymmetric context that could bias the model's decision toward that VL. If SWOT Analysis information were included in the above example, the model might weight VL\_B higher simply because additional contextual grounding was provided for that lever-not because VL\_B is genuinely more applicable to the scenario. By restricting context to shared tools, both VLs receive equivalent contextual support, ensuring the comparison reflects relative VL merit rather than differential context availability;
- Bias Mitigation Mechanism: This approach prevents localized impact where tool-specific terminology or recommendations in the prompt could activate favorable associations for one VL over another. The model evaluates both VLs against the same analytical framework, eliminating context asymmetry as a confounding factor in pairwise decisions;
- Shared Tools: When both VLs reference the same strategic tool, that tool's complete framework is included in the prompt context to enable informed cross-referencing and ensure consistency across the comparison;
- Empty Intersection Handling: When two VLs share no common tool references, the comparison proceeds with scenario information and VL metadata only, without strategic tool context. This represents a minority of comparisons given the three-tool ecosystem's coverage of P2P-related VLs.

This selective context injection mirrors RAG principles by retrieving and incorporating only relevant information, with an additional constraint requiring bilateral relevance to both comparison subjects. The approach ensures that ranking decisions are grounded in shared analytical frameworks while preventing context-induced bias that could distort pairwise preference signals. This design reflects a deliberate trade-off: accepting reduced context richness for individual comparisons in exchange for systematic fairness across the full pairwise comparison matrix.

5) LLM based solution Llama-3.1-8B-instruct numeric values with VL metadata (Experiment v3.1);

This configuration mirrors Experiment v2.1 while utilizing the larger model Llama-3.1-8B-instruct.

6) LLM based solution Llama-3.1-8B-instruct categorical values with VL metadata calibrated prompt (Experiment v3.2);

This configuration mirrors Experiment v2.2 while utilizing the larger model Llama-3.1-8B-instruct.

7) LLM based solution Llama-3.1-8B-instruct numeric values with VL metadata calibrated prompt no VL values (Experiment v3.3);

This configuration mirrors Experiment v2.3 while utilizing the larger model Llama-3.1-8B-instruct.

Prompt example:

...

### First Value Lever with business objective profile:

'Value Lever Name': 'Virtual supplier meetings'

'Addressable Business Objective': ['Sustainability', 'Efficiency']

'Hint': 'Reducing carbon emissions by traveling less adds up over time'

### Second Value Lever with business objective profile:

'Value Lever Name': 'Waste reduction'

'Addressable Business Objective': ['Sustainability']

'Hint': 'Align cross-functionally and with suppliers to avoid any form of waste (resources, materials, services) along the value chain'

...

The key difference is that we removed part related to VL table because it was found to introduce model bias:

...

'Savings/Margin': Very Low

'Sustainability': High

'Innovation': Very Low

'Agility': Very Low

'Resilience/Securing supply': Very Low

'Quality': Very Low

'Growth': Very Low

'Regulatory': Very Low

'Diversity': Very Low

'Efficiency': High

'Time to market': Very Low

...

8) LLM based solution Llama-3.1-8B-instruct numeric values with VL metadata calibrated prompt no VL values and PwP tools (Experiment v3.4);

This configuration extends Experiment v3.3 by incorporating strategic tool information. All experiments names, versions and short descriptions can be found in Table 8.

Note: Within each model series (v2.x and v3.x), experiments follow a single-variable design where each configuration changes exactly one factor from its predecessor, enabling isolation of individual effects. Model size comparison (v2.x vs v3.x) constitutes a separate

controlled comparison with matched configurations (v2.1 to v3.1, v2.2 to v3.2). The v3.3 configuration was added after v2 series analysis revealed VL metadata-induced bias, explaining the structural difference between series.

Table 8. Full list of experiments with description, model and number

Experiment version	Short description	Model	Change from Previous
v1	Original ranking with proprietary solution	Proprietary	Baseline
v2.1	numeric values with VL metadata	Llama-3.2-3B-instruct	LLM implementation (baseline for 3B series)
v2.2	categoric (calibrated) values with VL metadata	Llama-3.2-3B-instruct	Change numeric to categorical values in prompt
v2.3	categoric (calibrated) values with VL metadata with cross tools information	Llama-3.2-3B-instruct	Added strategic tools context
v3.1	numeric values with VL metadata	Llama-3.1-8B-instruct	model size: 3B → 8B (baseline for 8B series)
v3.2	categoric (calibrated) values with VL metadata	Llama-3.1-8B-instruct	Change numeric to categorical values in prompt
v3.3	categoric (calibrated) values without VL metadata	Llama-3.1-8B-instruct	Delete VL metadata (bias mitigation)
v3.4	categoric (calibrated) values without VL metadata but with tools information	Llama-3.1-8B-instruct	Added strategic tools context

## 4.2 Sample Size Justification

Sample size determination required balancing three competing factors: computational feasibility, statistical power, and input distribution requirements.

1) Computational Constraints: With 15 VLs, each experiment requires 105 non-repeating pair permutations. Single experimental runs require 2–6 hours of GPU processing time. With 8 experiments × 965 samples × 105 comparisons, the complete experimental pipeline required approximately 810600 individual LLM inference calls, totaling 40–80 GPU-hours. This aggregate computational overhead, combined with iterative prompt engineering requiring multiple full pipeline re-runs, establishes practical upper bounds on feasible sample sizes;

2) Input Distribution Requirements: The experimental design incorporates 11 scenarios with values ranging from 1 to 10. Stratified sampling across scenarios and values was used to approximate normal distribution and mitigate input bias. This approach ensures balanced representation across the parameter space while maintaining statistical validity;

3) Effect Size Considerations: The study aims to detect small to medium effects in VL ranking patterns. Chi-square goodness-of-fit testing was selected as the primary analytical framework due to its suitability for categorical ranking data. This nonparametric approach evaluates whether observed ranking frequencies deviate from expected distributions without imposing strict distributional assumptions;

Effect size quantification utilizes Cramér's  $V$ , which standardizes deviation magnitude independently of sample size. This metric enables comparison across studies and facilitates sample size estimation for detecting ranking patterns of varying strengths. Combined with noncentral chi-square methods, Cramér's  $V$  provides the foundation for determining minimum sample requirements at specified power levels.

#### Power Analysis Parameters:

- Alpha ( $\alpha$ ) = 0.05: Standard significance threshold, controlling Type I error probability;
- Target Power ( $1-\beta$ ) = 0.80: Ensuring 80% probability of detecting true effects when present;
- Degrees of Freedom = 14: Based on 15 ranking positions minus one;
- Effect Sizes (Cramér's V): Range 0.1-0.6, covering small to large effects.

4) Decision Framework: Samples exceeding 1,000 observations introduce prohibitive computational time durations (more than 6 hours per experiment) given LLM processing requirements and GPU hardware. At the same time, samples below 500 observations achieve adequate power ( $\geq 0.80$ ) only for effect sizes above Cramér's V = 0.2, risking failure to detect subtle yet meaningful ranking patterns given LLM output sensitivity;

5) Final Sample Configuration: Stratified sampling generated 965 observations, falling within the optimal 500-1,000 range. This configuration achieves adequate statistical power for detecting small-to-medium effects while maintaining computational tractability and satisfying distributional requirements across experimental parameters as can be observed in Fig 2.

The chi-square goodness-of-fit framework and Cramér's V effect sizes serve dual analytical purposes in this study. First, they provide the statistical foundation for sample size determination as described above. Second, they serve as complementary distributional metrics reported alongside Kendall's  $\tau$  in Section 5, where chi-square tests evaluate whether observed VL ranking frequencies deviate significantly from uniform distributions, and Cramér's V quantifies the magnitude of these deviations.

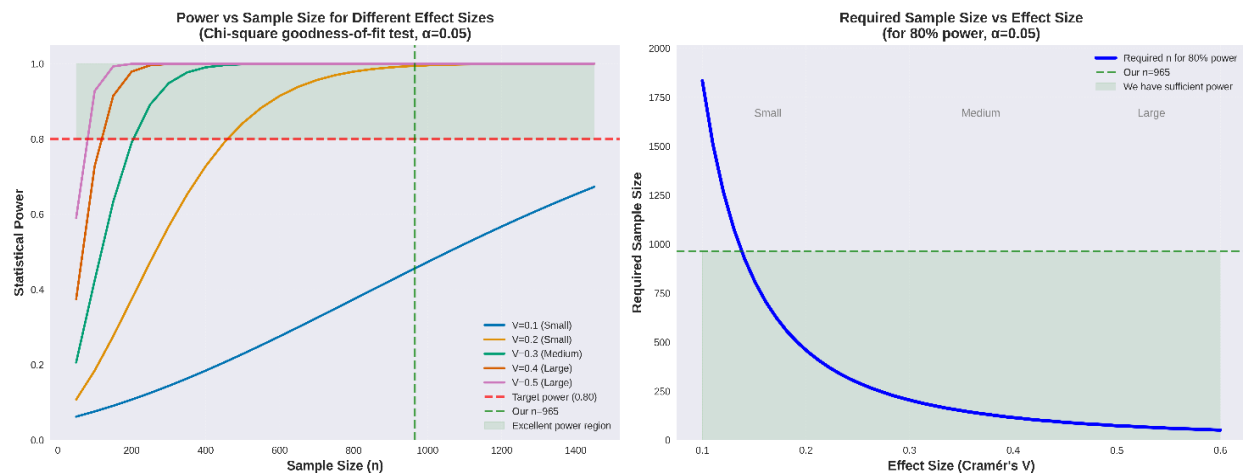


Fig 2. Power vs sample size for different effect sizes (left) and required sample size vs effect size (right)

This dual-framework approach captures both ordinal association (Kendall's  $\tau$ : how well do LLM rankings correlate with baseline) and distributional structure (chi-square/Cramér's V: how deterministic are individual VL positions). The two metrics address distinct but complementary research questions.

## 5.0 Results and Analysis

### 5.1 Comparing of v1, v2.1, v2.2 and v2.3 experiments

Kendall's Tau and Weighted Kendall's Tau. First metrics that we are going to use is Kendall's Tau ( $\tau$ ) and Weighted Kendall's Tau, we are going to compare v1 to v2.1, v1 to v2.2 and v1 to 2.3

Kendall's Tau quantifies the ordinal association between two ranked lists by evaluating concordant versus discordant pairs of items. The metric ranges from  $-1$  (complete rank reversal) through  $0$  (no systematic association) to  $+1$  (perfect rank agreement).

**Weighted Kendall's Tau Extension.** The weighted variant of Kendall's Tau introduces position-dependent importance factors, assigning greater significance to disagreements among top-ranked items relative to lower-ranked positions. This modification reflects practical scenarios where ranking accuracy at the list head carries disproportionate value as observed in search result relevance, and priority-driven decision frameworks like VL in procurement scenarios due to need to pick only few VL to work with and leverage their value. While the theoretical range remains analogous to standard Tau (approximately  $-1$  to  $+1$ ), the weighting scheme and tie-handling mechanisms may introduce implementation-specific interpretation considerations.

Based on the chart that is visible on Fig. 3, we can see weak to medium correlation (0.358 for basic v2.1 and higher 0.381 for v2.3 extended with tools experiment) with v1 experiment, interestingly when we take a look at non weighted version it also increases in correlation with v1 more, due to fact that tailing ranks are close to the original predictions.

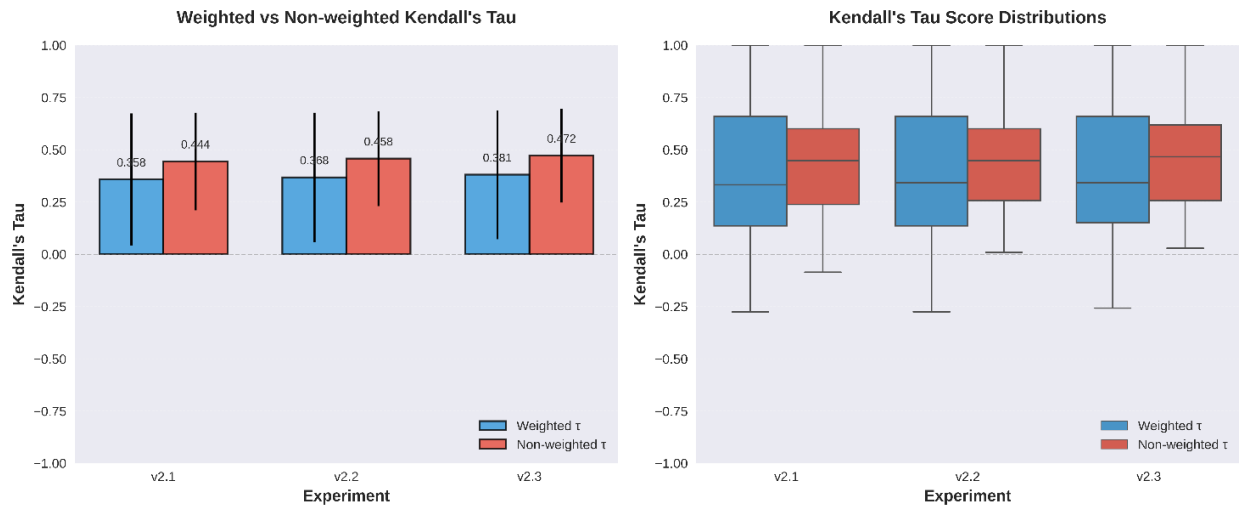


Fig 3. Weighted and non-weighted Kendall's Tau with respect to v1 and weighted and non-weighted Kendall's Tau score distributions.

**Intra-Experiment Ranking Consistency.** Beyond comparing LLM-generated rankings against baseline, we evaluated ranking stability within each experiment. This analysis quantifies how consistently the same experimental configuration produces similar rankings across different input samples, independent of ground truth accuracy.

Two metrics captured internal consistency:

- 1) **Pairwise Displacement:** Average absolute position difference computed across all ranking pairs within an experiment. Lower values indicate more reproducible ranking behavior under identical experimental conditions;
- 2) **Position Variance:** Variance in ranking positions for each PwP VL across all samples in an experiment. Lower variance indicates stable item positioning regardless of specific input combinations;

Lower displacement and variance values indicate more deterministic ranking behavior, where the solution produces consistent outputs given similar inputs. Higher values suggest sensitivity to input variations or stochastic elements in the ranking process. This internal consistency metric complements accuracy measures: an experiment may achieve high baseline agreement while exhibiting low consistency, or conversely, maintain stable but systematically biased rankings.

Analysis of intra-experiment consistency metrics in Fig. 4 combined with PwP VL ranking frequency patterns from Figures 4,5,6,7,8 reveals significant determinism in LLM-based ranking approaches.

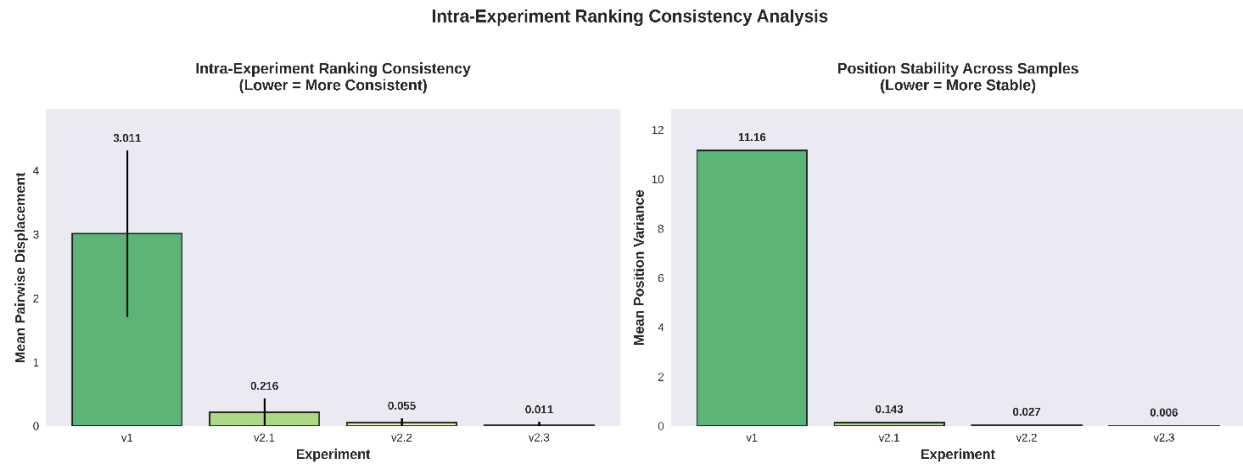


Fig 4. Intra-Experiment Ranking Consistency Analysis

The observed consistency levels suggest overconfident model behavior, potentially attributable to the internal voting mechanism used in the ranking process. However, the magnitude of consistency exceeds expected levels, indicating potential model or data bias.

Comparative analysis of experiments v2.1 and v2.2 demonstrates the impact of prompt formulation on ranking stability. These experiments differ only in value representation format: numeric (1-9 scale) versus categorical descriptors ("Very Low" to "Critical"), yet position stability improved substantially from 0.143 to 0.027.

This shift indicates that minor prompt modifications produce significant changes in output rankings, revealing that LLM internal representations of numeric and categorical values are distinct enough to take into account. This finding has practical implications for prompt engineering in VL ranking applications or any prompts incorporating or working with any kind of numeric and categorical values. The introduction of strategic tool information in experiment v2.3 further increased ranking determinism, as reflected in reduced position variance. While this enhancement improves consistency, it simultaneously amplifies the bias concern identified in earlier experiments.

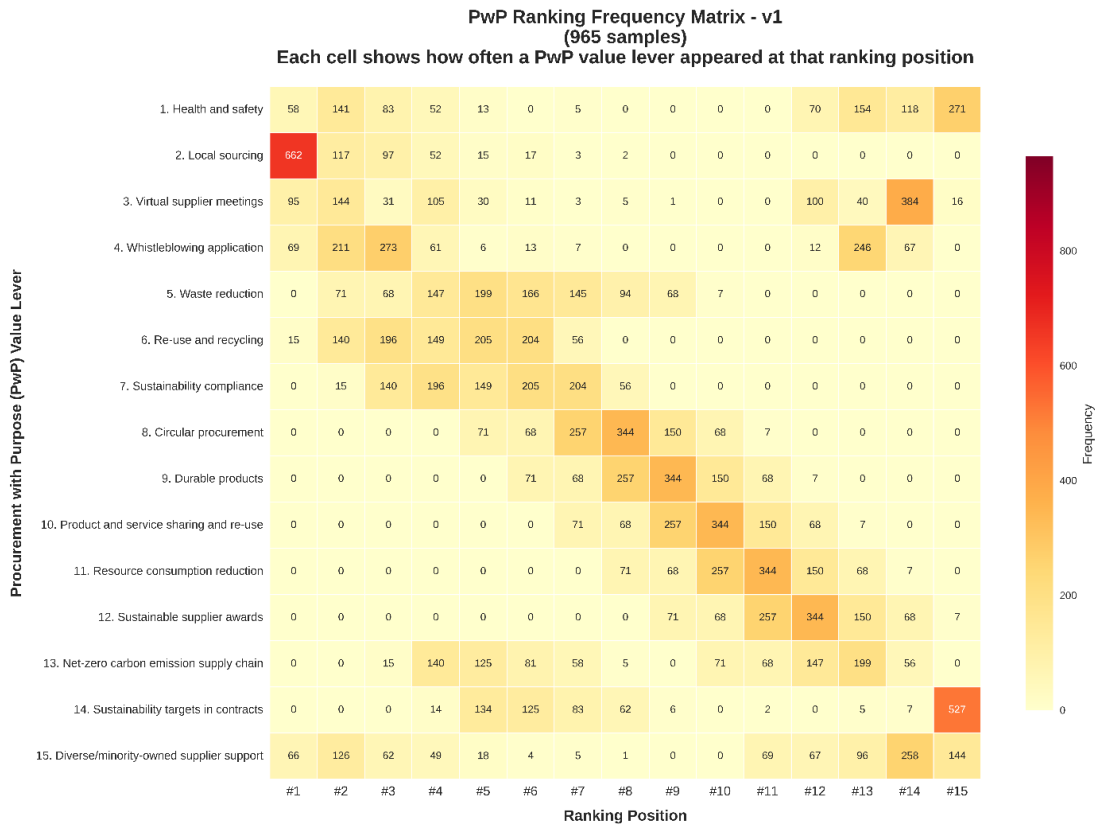


Fig 5. PwP VLs Ranking frequency matrix for experiment v1

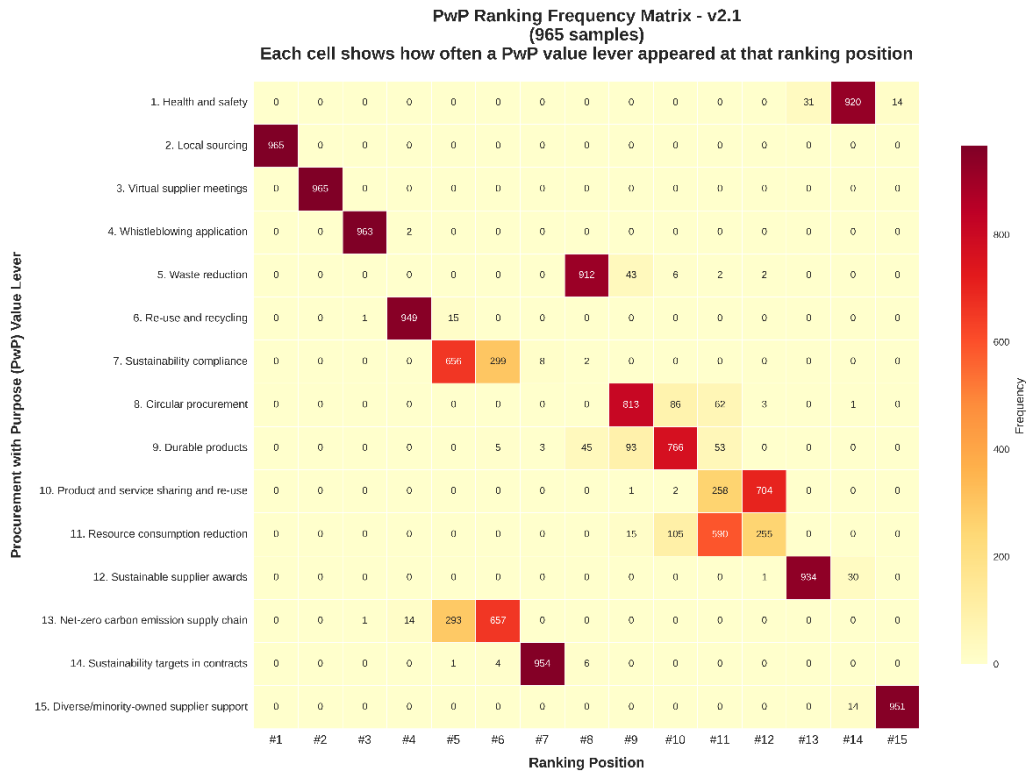


Fig 6. PwP VLs Ranking frequency matrix for experiment v2.1

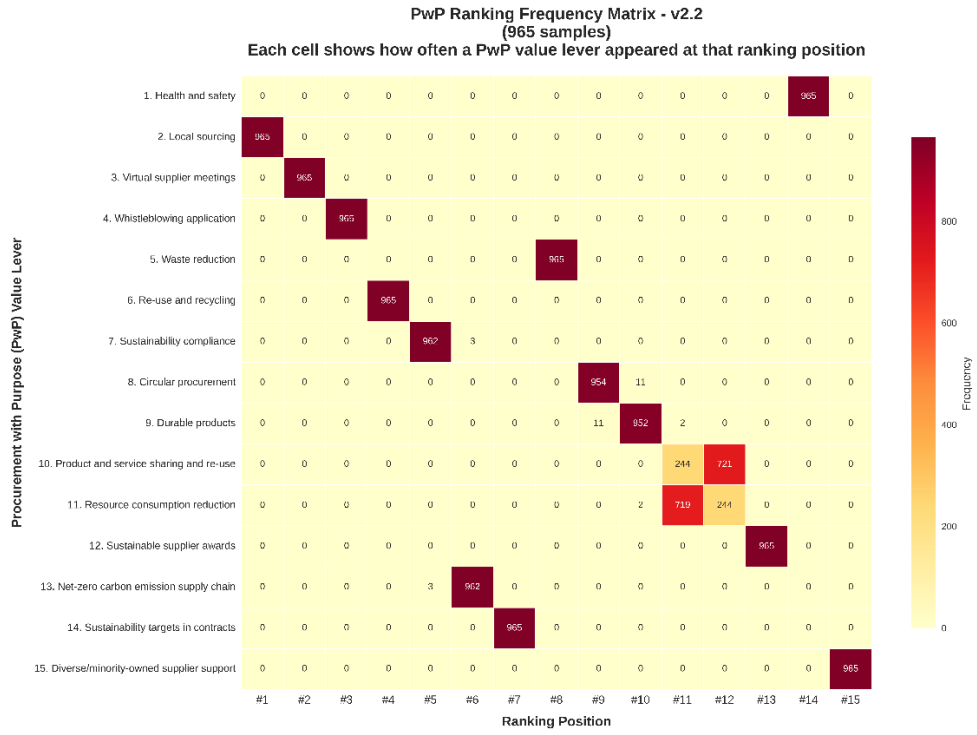


Fig 7. PwP VLs Ranking frequency matrix for experiment v2.2

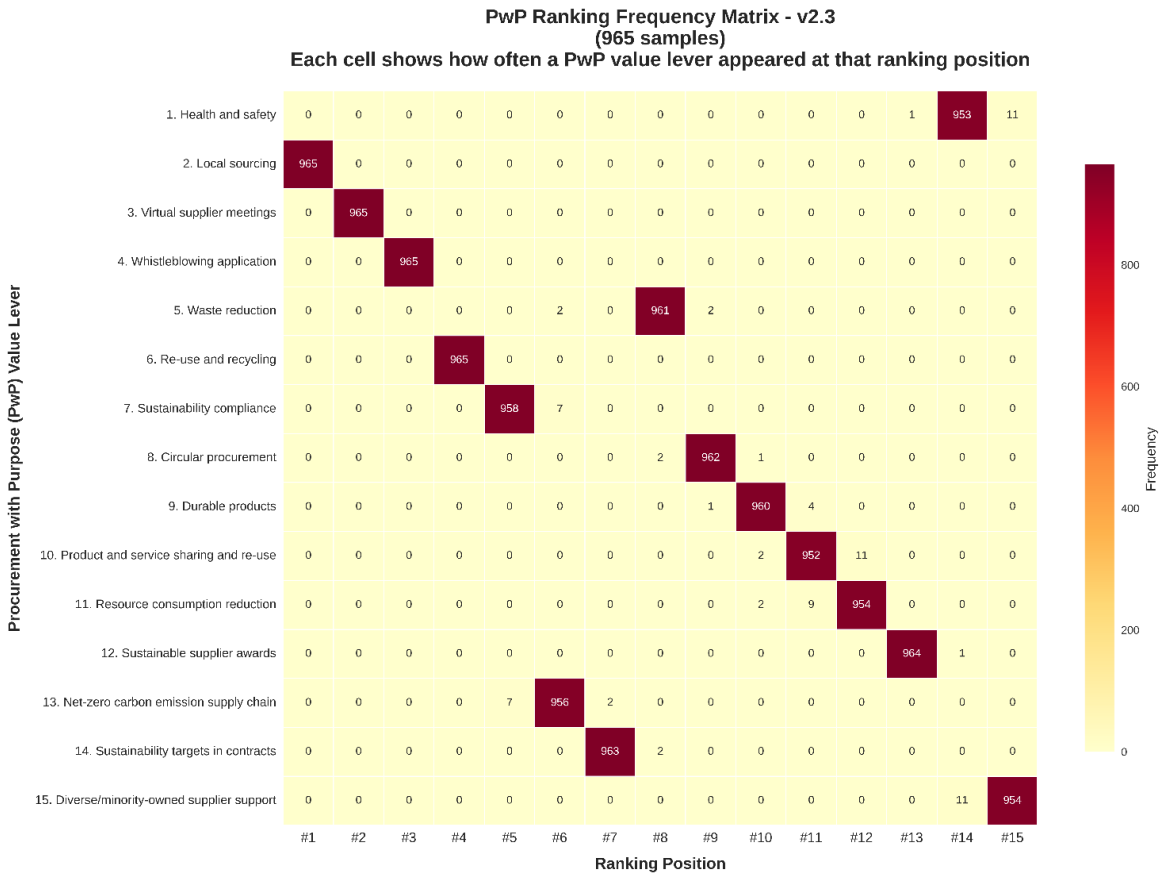


Fig 8. PwP VLs Ranking frequency matrix for experiment v2.3

The progressive reduction in ranking variability across v2.1 → v2.2 → v2.3 suggests systematic increase in model confidence that may not reflect genuine domain understanding but rather reinforcement of existing representational patterns. These results highlight a critical tension in LLM-based ranking systems: increased consistency does not necessarily indicate improved accuracy or domain alignment. The dramatic stability improvements observed through relatively minor prompt modifications suggest that rankings may reflect learned patterns in training data rather than robust procurement domain reasoning. Future development of such systems must account for these representational differences and implement validation mechanisms that distinguish between genuine consistency and overconfident determinism. Same finding we observed during case study variance for particular VL that can be seen on Fig. 9.

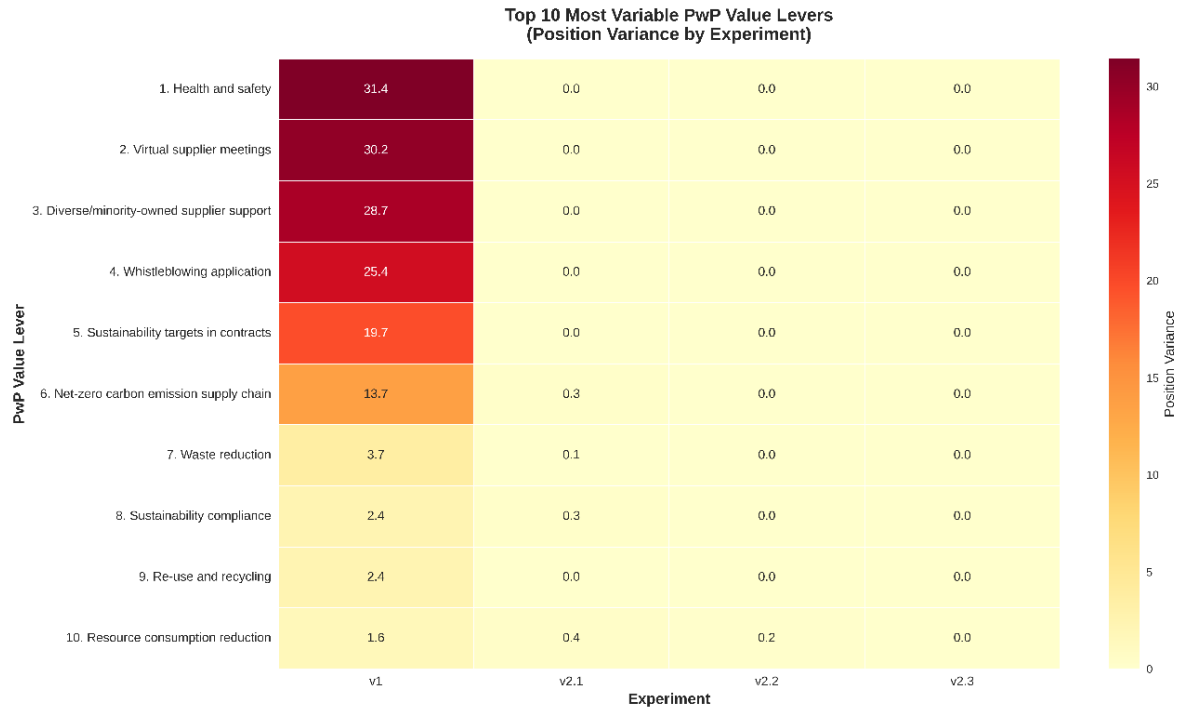


Fig 9. Case study of top 10 most variable VLs for experiments v1, v2.1, v2.2 and v2.3

## 5.2 Comparing of v1, v3.1, v3.2, v3.3 and v3.4 experiments

The v3 experimental series uses Llama-3.1-8B-instruct, a larger model compared to the Llama-3.2-3B-instruct variant used in v2 experiments. This increased model capacity enables more precise outputs and improved reasoning capabilities. Additionally, v3 experiments incorporate design modifications to address bias patterns identified in the v2 experimental series, including refinements to prompt structure and context injection mechanisms. As anticipated, experiments v3.1 and v3.2 exhibited persistent bias patterns and moderate correlation with the v1 baseline that can be seen on Fig. 10.

However, experiment v3.3, which removed internal VL numeric data from prompts, weighted Kendall's tau decreased approximately twofold compared to v3.1 and v3.2. This reduction indicates weaker alignment with baseline rankings, suggesting the model's ranking decisions became less deterministic and potentially less biased toward the input data.

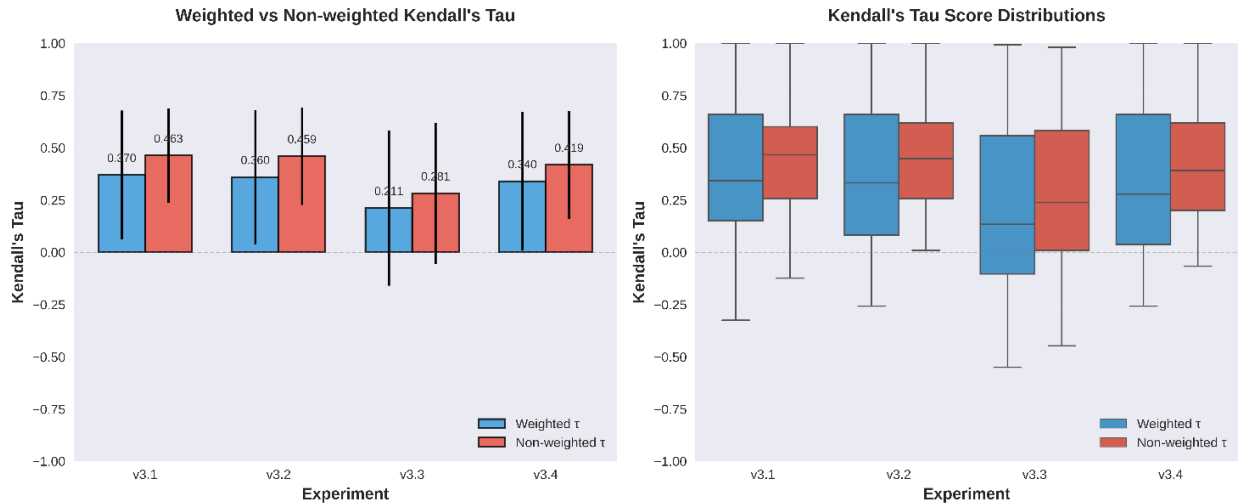


Fig 10. Weighted and non-weighted Kendall's Tau with respect to v1 and weighted and non-weighted Kendall's Tau score distributions.

With intra-experiment consistency metrics shown in Fig. 11: the v3.3 experiment demonstrated position stability of 9.348, substantially closer to v1 baseline levels compared to other v3 experiments. This alignment indicates that removing internal VL numeric data reduced deterministic bias while maintaining comparable consistency patterns to the proprietary baseline.

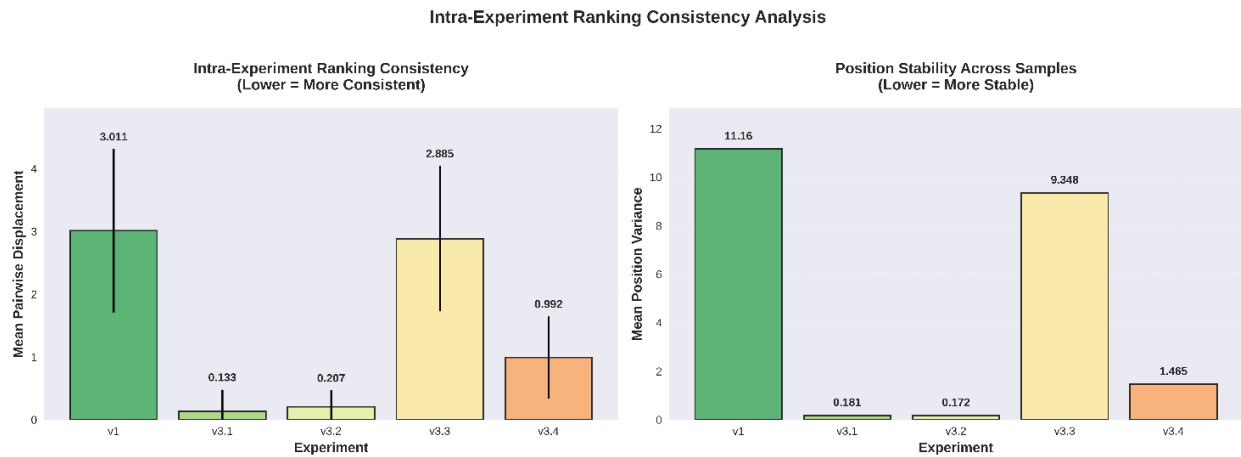


Fig. 11. Intra-Experiment Ranking Consistency Analysis

This is also clearly visible in Figures 12, 13 versus Figures 14, 15.

The subsequent addition of strategic tool context in v3.4 dramatically increased stability, reducing mean pairwise displacement from 9.348 to 1.465 a 6.38-fold improvement. This substantial shift demonstrates the model's capacity to incorporate structured contextual information from procurement frameworks (KM, PwP Framework, SWOT Analysis) into ranking decisions, producing highly deterministic outputs.

However, this increased determinism introduces a critical distinction from the v1 baseline: while v1 exhibits moderate consistency reflecting expert judgment variability, v3.4's low position variance (1.465) suggests potential over-fitting to provided tool context. The position stability metric reveals that v3.4 rankings become highly predictable given input configuration - a characteristic absent in v1's approach, which maintains flexibility across diverse procurement scenarios.

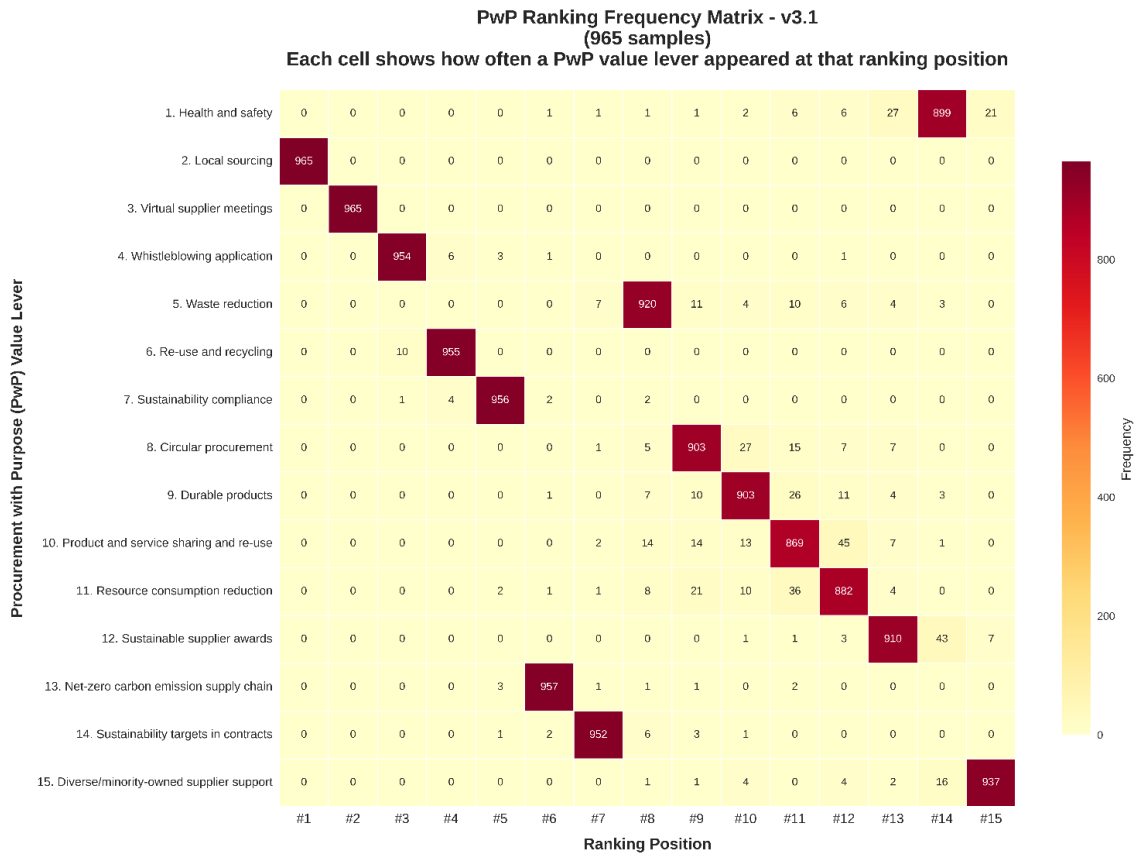


Fig 12. PwP VLs Ranking frequency matrix for experiment v3.1

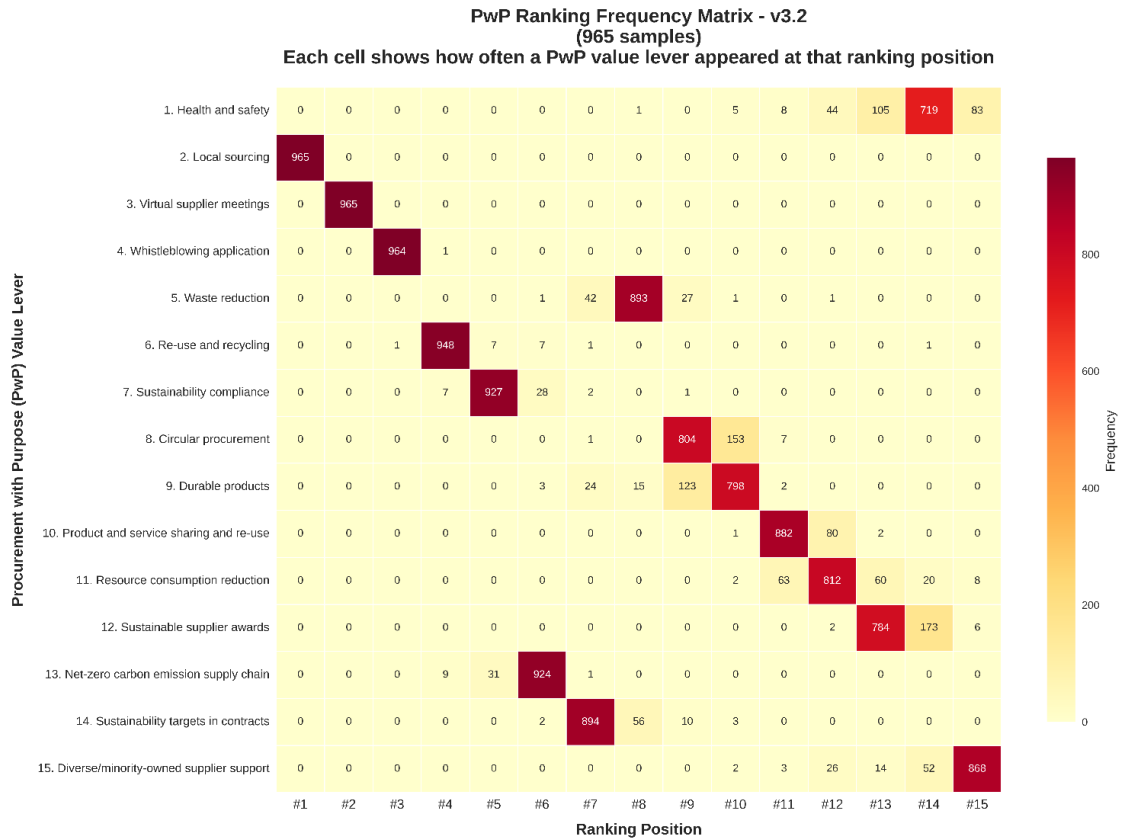


Fig 13. PwP VLs Ranking frequency matrix for experiment v3.2

This divergence highlights a fundamental limitation of the proprietary v1 solution: its inability to systematically integrate additional contextual frameworks while preserving ranking flexibility. The LLM based approach demonstrates capability for context incorporation, though at the cost of introducing deterministic bias patterns that require careful management in operational deployment.

Additionally, if we compare v2.1 vs v2.2 and v3.1 vs 3.2, v2.1 vs v2.2 with a smaller 3b model showed high change on stability due to prompt calibration, while the 8b model tends to be more resilient to these factors, most likely due to better encoding / decoding capabilities.

Evidence of Model/Voting Induced Bias: Analysis of individual PwP VL variance across experiments (Figure 16) demonstrates distributional changes affecting a minority of VLs (2-3 of 15, representing 13-20% of the portfolio). Two representative cases illustrate this phenomenon:

- Virtual Supplier Meetings: Position variance decreased from 30.2 (v1) to 0.2 (v3.3) a 150-fold reduction. This stabilization indicates strong model preference for specific ranking positions. However, this pattern admits multiple interpretations beyond pure model bias. From a procurement domain perspective, certain VLs represent near-universal best practices that experienced practitioners consistently prioritize regardless of scenario variation. "Virtual Supplier Meetings" exemplifies such a lever: reducing travel-related carbon emissions and costs applies broadly across procurement contexts, potentially justifying consistent high ranking. The model may be capturing genuine domain knowledge rather than exhibiting arbitrary bias;
- Sustainability Compliance: Position variance increased from 2.4 (v1) to 22.4 (v3.4) a 9.3-fold increase. This shift demonstrates the inverse pattern, where LLM based approaches introduce ranking instability absent in the baseline solution. The elevated variance suggests the model lacks consistent internal representation for this VL, leading to position fluctuations across similar input configurations.

These contrasting patterns affect a minority of VLs while the remaining 80-87% demonstrate appropriate scenario responsiveness. Critically, the specific VL names likely influence model behavior substantially: LLMs encode semantic associations from pre-training corpora that may systematically favor or disfavor particular terminology. Future research should investigate how VL naming conventions affect ranking outputs through controlled experiments varying VL labels while holding underlying definitions constant.

Additionally, the pairwise voting aggregation mechanism likely contributes to these per-VL artifacts. No aggregation algorithm is perfect, and the binary vote accumulation process amplifies small preference margins into pronounced ranking differences.

The ranking frequency heatmaps corroborate these findings, showing systematic concentration of specific VLs in particular position ranges for LLM experiments compared to the more dispersed distributions observed in v1.

This concentration pattern indicates that prompt engineering modifications (v3.1 → v3.2 → v3.3 → v3.4) successfully reduced overall bias metrics but simultaneously introduced position-specific structural artifacts that manifest as extreme variance changes individual VLs.

The observed ranking determinism stems from two distinct mechanisms. First, LLM internal representations impose structural constraints on VL positioning, as evidenced by extreme variance shifts (e.g., Virtual Supplier Meetings: 30.2 → 0.2). Second, the pairwise voting aggregation mechanism introduces additional distributional effects analogous to softmax normalization.

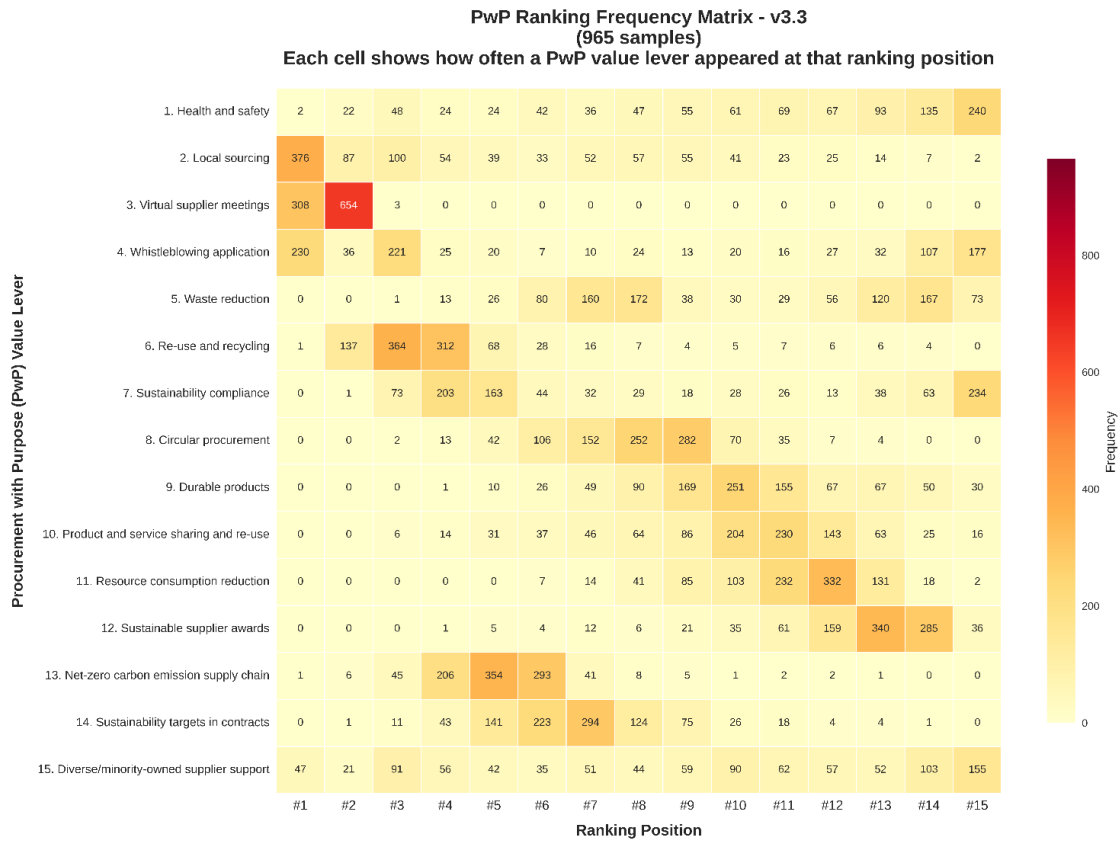


Fig 14. PwP VLs Ranking frequency matrix for experiment v3.3

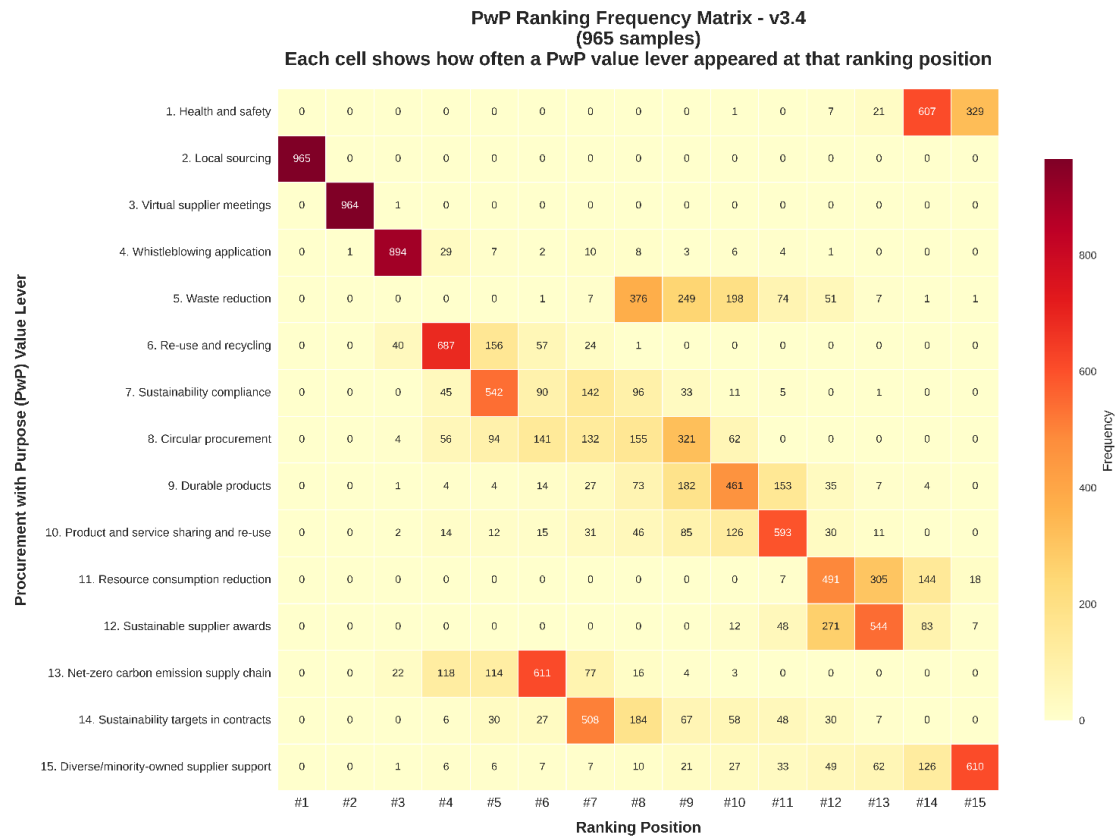


Fig 15. PwP VLs Ranking frequency matrix for experiment v3.4

The voting process operates across multiple pairwise comparisons, where each VL accumulates votes based on binary preferences. This aggregation mechanism amplifies consensus patterns: VLs receiving slight preference majorities in individual comparisons gain disproportionate representation in final rankings. The effect mirrors softmax behavior: small differences in raw preference scores become magnified through the voting accumulation process, producing sharper, more deterministic final distributions.

This dual mechanism (model bias + voting amplification) explains why v3 experiments achieve high internal consistency (Section 4.B: v3.4 displacement = 1.465). The combination transforms probabilistic LLM outputs into highly predictable rankings that may not reflect genuine domain reasoning but rather systematic reinforcement of learned patterns.



Fig 16. Case study of top 10 most variable VLs for experiments v1, v3.1, v3.2, v3.3 and v3.4

The trade-off between consistency and flexibility becomes apparent: v3 experiments produce reliable, reproducible rankings suitable for systematic expert review, but lack the adaptive variability characteristic of expert decision-making across diverse procurement scenarios. The voting mechanism, while enabling consensus aggregation, may inadvertently suppress minority preferences that could be relevant in specific contexts.

Scope Limitation and Category Effects: This study focused exclusively on PwP-related VLs a single procurement category comprising 15 items. Operational procurement portfolios typically span 5-6 distinct categories (e.g., commercial, technical, strategic, compliance, sustainability, innovation) with 50+ total VLs exhibiting different semantic characteristics and domain associations. The per-VL variance patterns documented here may be category-specific: PwP VLs share sustainability-focused terminology that could trigger consistent LLM associations. Expanding analysis to multi-category portfolios would likely reveal different variance distributions, potentially with fewer extreme cases as category diversity dilutes any single semantic pattern.

This analysis reveals that the automated ranking system achieves operational utility through high consistency and predictability, attributes valuable for expert validation workflows. However, the underlying determinism suggests that rankings reflect systematic pattern application rather than dynamic contextual reasoning, a limitation that

must be acknowledged in deployment scenarios requiring adaptability to novel or edge-case procurement situations.

### 5.3 Bootstrap Validation of Ranking Correlations

To strengthen the empirical grounding of Kendall's  $\tau$  estimates, bootstrap resampling ( $n=5,000$  iterations) was performed to compute 95% confidence intervals for all experimental configurations. Table 9 presents the results. LLM inference was conducted with temperature=0 (greedy decoding), ensuring deterministic outputs that eliminate seed-dependent variability in pairwise comparisons.

Table 9: Bootstrap 95% CI for Kendall's  $\tau$  and Weighted Kendall's  $\tau$

Experiment version	$\tau$	95% CI	$\tau_w$	95% CI
v2.1	0.444	[0.429, 0.459]	0.498	[0.485, 0.511]
v2.2	0.458	[0.443, 0.472]	0.511	[0.499, 0.523]
v2.3	0.472	[0.458, 0.486]	0.520	[0.507, 0.532]
v3.1	0.463	[0.449, 0.477]	0.514	[0.502, 0.526]
v3.2	0.459	[0.445, 0.474]	0.511	[0.499, 0.524]
v3.3	0.281	[0.260, 0.303]	0.332	[0.312, 0.352]
v3.4	0.419	[0.402, 0.435]	0.472	[0.458, 0.487]

The narrow confidence intervals ( $\pm 0.015$  on average) indicate statistically robust correlation estimates across all configurations. Notably, v3.3 exhibits significantly lower correlation [0.260, 0.303] compared to other experiments, with non-overlapping confidence intervals confirming the statistical significance of the bias reduction achieved through VL metadata removal. The consistent difference between weighted and non-weighted  $\tau$  (non-overlapping CIs) validates that top-position ranking agreement is systematically stronger than overall ordinal correlation as shown in Fig. 17.

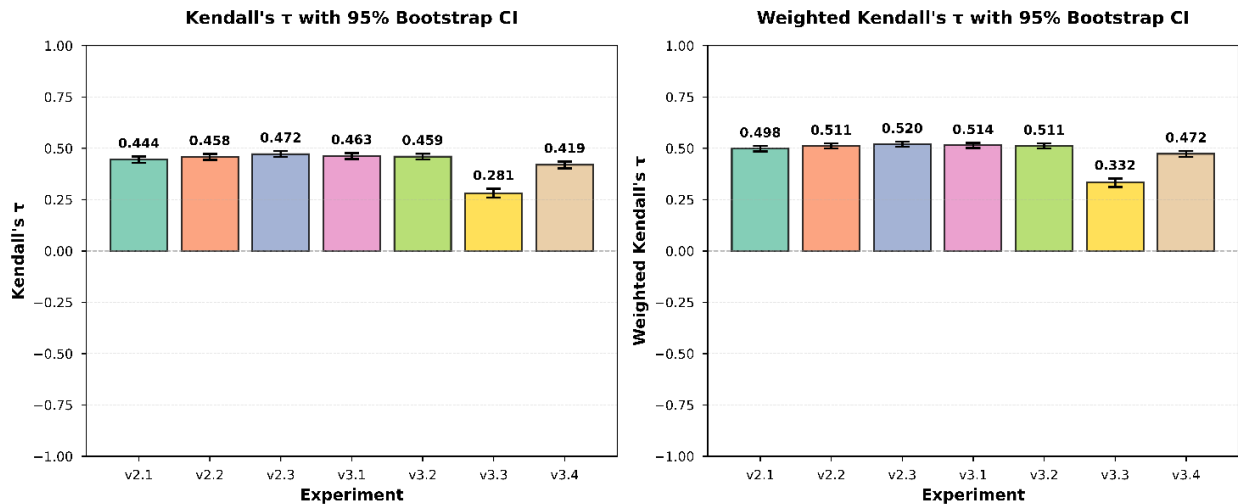


Fig 17: Kendall's  $\tau$  with 95% Bootstrap CI (left) and Weighted Kendall's  $\tau$  with 95% Bootstrap CI

### 5.4 Variance Reduction Analysis and Expert Review Implications

To quantify the practical impact of strategic tool integration on ranking stability, we analyze position variance reduction across the full experimental progression, measured over all 965 scenario configurations per experiment.

Experiments v2.1–v3.2, which include VL numeric metadata in prompts, exhibit artificially low position stability (0.027–0.143), reflecting bias-induced determinism rather than genuine

ranking quality (Section 5.1). These configurations are excluded from variance reduction calculations as their stability is inflated by input data leakage.

Experiment v3.3 establishes the unbiased LLM baseline: after removing VL metadata to eliminate identified bias sources, position stability rises to 9.348, comparable to the v1 proprietary baseline variability.

This configuration represents the expected variance when the model ranks VLs based solely on scenario context, VL names, addressable business objectives, and implementation hints that are without accessing to answer-correlated numeric profiles or strategic tool outputs.

The subsequent addition of strategic tool context (KM, PwP Framework, SWOT Analysis) in v3.4 reduces position stability from 9.348 to 1.465, representing an 84.3% variance reduction computed as  $(9.348 - 1.465) / 9.348$ . This reduction is measured across all 965 scenario configurations, establishing it as a multi-scenario validated result rather than a single-case observation. Experiment v3.4 represents the strongest configuration in this study: it achieves stability improvements through legitimate contextual grounding while structured strategic frameworks providing domain-relevant reasoning signals without data leakage or bias amplification observed in earlier configurations. While correlation with the v1 baseline remains moderate (weighted  $\tau \approx 0.38$ ), v3.4 uniquely combines three properties: (1) absence of identified input bias sources, (2) high ranking consistency across scenarios, and (3) contextual grounding in established procurement frameworks, making it the most operationally viable configuration for deployment.

Implications for Expert Review Scope. Position variance directly determines how many VL positions require substantive expert assessment versus cursory validation in operational deployment. Under v3.3, high positional variance means most of the 15 VL positions fluctuate meaningfully across scenarios, requiring expert deliberation comparable to fully manual prioritization. Under v3.4, the 84% variance reduction concentrates instability in a small subset of positions. Thus, experts need only perform deep review on the few remaining high-variance VLs while rapidly confirming stable positions.

We emphasize that this 84% figure quantifies variance reduction in ranking output, not a directly measured reduction in expert hours. The translation from reduced variance to reduced workload depends on organizational review processes, reviewer experience, and procurement context specifics. Nevertheless, the mechanism is direct: fewer unstable positions require fewer expert deliberation cycles. Combined with the sixfold stability improvement documented in Section 5.2, these results establish a quantitative basis for reduced expert involvement in the ranking validation process when strategic tool context is available.

### 5.5 Chi-Square Distributional Analysis and Power Validation

To complement the ordinal correlation analysis presented in Sections 5.1 and 5.2, we report chi-square goodness-of-fit results and Cramér's  $V$  effect sizes as specified in the sample size justification framework (Section 3. 7). While Kendall's  $\tau$  captures pairwise ordinal agreement between experimental rankings and the v1 baseline, chi-square testing evaluates a fundamentally different property: whether the observed ranking position frequencies for each VL deviate significantly from a uniform distribution across the 15 positions. This distributional perspective quantifies the degree of positional concentration, the extent to which a VL is constrained to specific ranking positions rather than appearing with equal probability across all positions.

For each of the 15 VLs within each experiment, a chi-square goodness-of-fit test was conducted against the null hypothesis of uniform position occupancy (expected frequency =  $N/15$ , where  $N = 965$ ). Cramér's  $V$  was computed as the standardized effect size measure, enabling direct comparison of distributional strength across VLs and experiments independently of sample size.

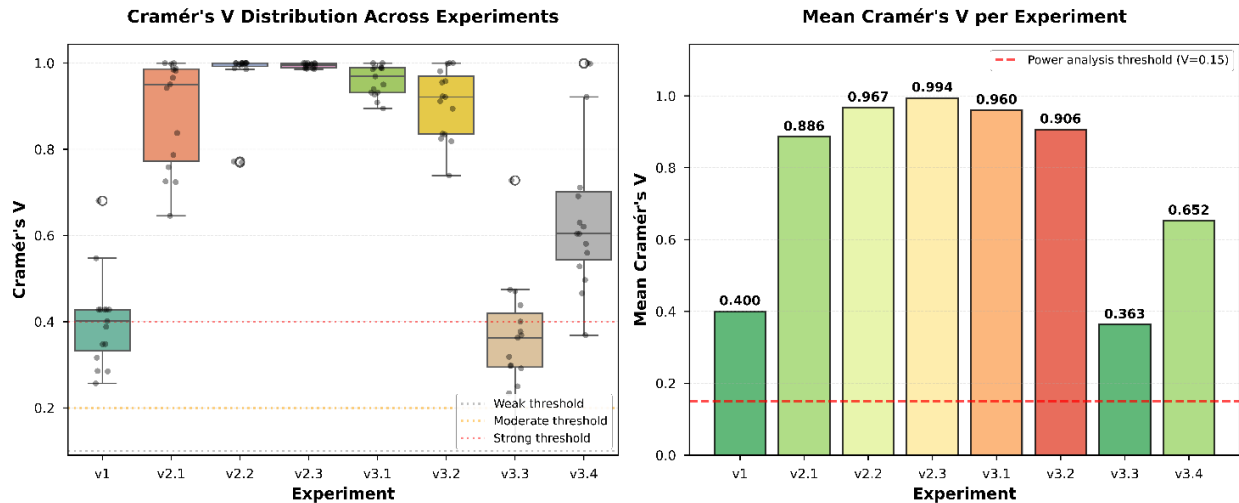


Fig 18. Cramér's V Distribution across experiments (left) and Mean Cramér's V per experiment

Proprietary Baseline (v1). Chi-square tests reject the uniform distribution hypothesis for all 15 VLs ( $p < 0.05$ ), with Cramér's V values ranging from 0.257 to 0.681 (mean  $V = 0.400$ ). This indicates strong distributional structure: v1 rankings exhibit systematic positional preferences rather than random assignment, yet the moderate mean V reflects the variability characteristic of expert-driven ranking systems that adapt outputs to diverse input configurations. The highest effect size ( $V = 0.681$ , Local Sourcing) indicates strong positional concentration, while the lowest ( $V = 0.257$ , Net-Zero Carbon Emission Supply Chain) demonstrates broader distributional dispersion, consistent with the position variance patterns documented in Section 5.1.

Biased Configurations (v2.1–v3.2). Under configurations incorporating VL metadata (v2.1, v2.2, v2.3, v3.1, v3.2), chi-square tests reject the uniform distribution hypothesis for all 15 VLs ( $p < 0.001$ ) in every configuration, with Cramér's V values ranging from 0.646 to 1.000 (cross-experiment mean  $V = 0.943$ ). These very strong effect sizes indicate extreme positional concentration consistent with the bias-induced determinism documented in Sections 4. A and 4. B. The near-ceiling V values corroborate the ranking frequency heatmap patterns (Figures 5, 6, 7, 11, 12), where specific VLs occupy narrow position ranges with minimal variability across the 965 input scenarios.

Unbiased Baseline (v3.3). Chi-square tests reject the uniform distribution hypothesis for all 15 VLs ( $p < 0.05$ ), with Cramér's V values ranging from 0.134 to 0.728 (mean  $V = 0.363$ ). Notably, v3.3 exhibits distributional characteristics closely aligned with v1: both experiments demonstrate moderate mean effect sizes (v1: 0.400, v3.3: 0.363), confirming that removing VL metadata from prompts successfully restores distributional patterns comparable to the expert-driven proprietary baseline. The slightly lower mean V for v3.3 reflects the broader positional dispersion visible in Figure 18, where rankings exhibit greater flexibility across input configurations relative to metadata-augmented experiments. The highest individual V (0.728, Virtual Supplier Meetings) indicates that this VL retains strong positional preferences even without metadata, consistent with the extreme variance reduction documented in Section 5.2 (30.2  $\rightarrow$  0.2).

Tool-Augmented Configuration (v3.4). Chi-square tests reject the uniform distribution hypothesis for all 15 VLs ( $p < 0.05$ ), with Cramér's V values ranging from 0.369 to 1.000 (mean  $V = 0.652$ ). The mean effect size positions v3.4 between the unbiased v3.3 (0.363) and the metadata-biased configurations (0.943), indicating that strategic tool integration (KM, PwP Framework, SWOT Analysis) introduces substantial positional concentration without reaching the extreme determinism characteristic of metadata-driven bias. Two VLs achieve near-perfect concentration ( $V \geq 0.999$ : Local Sourcing, Virtual Supplier Meetings), while the

minimum V (0.369, Circular Procurement) indicates that tool context does not uniformly constrain all VLs, preserving partial distributional flexibility as shown in Fig. 19.

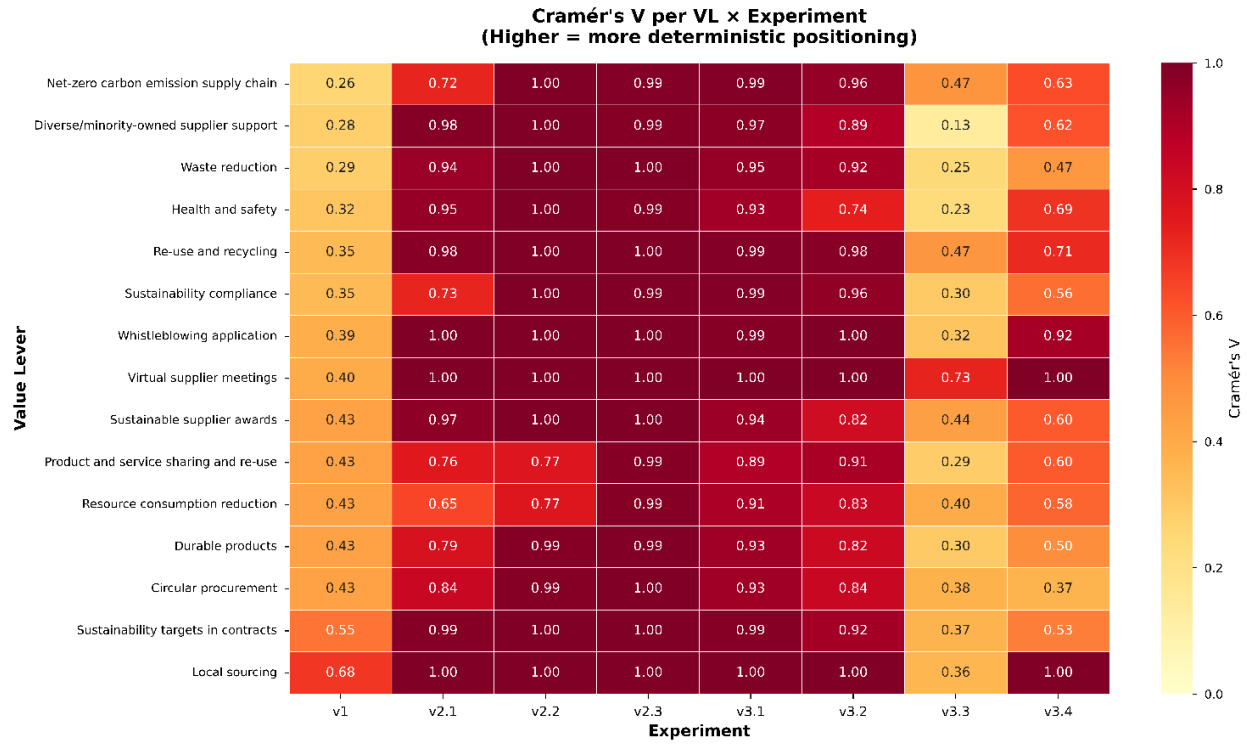


Fig 19. Cramér's V heatmap across 15 VL and 8 experiments

Power Validation. Post-hoc power analysis was conducted for all 120 VL-experiment pairs (15 VLs × 8 experiments) using the observed Cramér's V values, the critical chi-square threshold ( $\chi^2(14) = 23.68$  at  $\alpha = 0.05$ ), and the noncentral chi-square distribution. All 120 pairs achieve statistical power of 1.000, confirming that the sample size of N = 965 provides sufficient sensitivity to detect the observed distributional effects with certainty. This result validates the sample size determination framework established in Section 3. G: the selected sample configuration not only satisfies the target power threshold ( $1 - \beta \geq 0.80$ ) but substantially exceeds it, eliminating concerns regarding Type II error across all experimental conditions as shown in Table 10.

Table 10: Chi-Square Goodness-of-Fit and Cramér's V Summary Across Experiments  $H_0$ : Uniform distribution across 15 rank positions | df = 14 | Critical  $\chi^2(14, \alpha=0.05) = 23.68$

Experiment #	N	Sig. VLs (p<0.05)	Mean $\chi^2$	$\chi^2$ Range	Mean V	V Range	Mean Power	Min Power	Strength
v1	965	15/15	2306.0	[891.1, 6256.3]	0.400	[0.257, 0.681]	1	1	Moderate
v2.1	965	15/15	10803.3	[5631.5, 13510.0]	0.886	[0.646, 1.000]	1	1	Very Strong
v2.2	965	15/15	12718.4	[7996.2, 13510.0]	0.967	[0.769, 1.000]	1	1	Very Strong
v2.3	965	15/15	13355.4	[13124.6, 13510.0]	0.994	[0.986, 1.000]	1	1	Very Strong
v3.1	965	15/15	12460.8	[10814.3, 13510.0]	0.960	[0.895, 1.000]	1	1	Very Strong
v3.2	965	15/15	11180.9	[7380.6, 13510.0]	0.906	[0.739, 1.000]	1	1	Very Strong

Experiment #	N	Sig. VLS ( $p < 0.05$ )	Mean $\chi^2$	$\chi^2$ Range	Mean V	V Range	Mean Power	Min Power	Strength
v3.3	965	15/15	2014.3	[241.3, 7158.1]	0.363	[0.134, 0.728]	1	1	Moderate
v3.4	965	15/15	6189.1	[1836.1, 13510.0]	0.652	[0.369, 1.000]	1	1	Very Strong

Bridging Chi-Square and Kendall's  $\tau$ . The chi-square and Kendall's  $\tau$  metrics serve complementary analytical purposes. Kendall's  $\tau$  quantifies the ordinal agreement between LLM-generated rankings and the v1 baseline, measuring how consistently the relative ordering of VLS is preserved. Chi-square goodness-of-fit testing, by contrast, evaluates the internal distributional structure of each experiment independently, quantifying whether individual VLS exhibit systematic positional preferences. A configuration may achieve moderate Kendall's  $\tau$  (indicating imperfect ordinal agreement with baseline) while simultaneously exhibiting high Cramér's V (indicating strong internal positional structure). This pattern characterizes v3.4: moderate baseline correlation (weighted  $\tau = 0.45$ ) combined with strong distributional concentration (mean V = 0.652) indicates that tool-augmented rankings impose consistent positional structure that partially diverges from the proprietary baseline ordering. The dual-metric framework thus provides a more complete characterization of ranking behavior than either metric in isolation, distinguishing between ordinal alignment accuracy and positional consistency.

## 6.0 Discussions

### 6.1 Scalability Constraints and Adaptive Sampling Strategies

This study constrained analysis to 15 PwP-related VLS from operational portfolios typically exceeding 50 VLS. Scaling to 50+ VLS requires  $C(50,2) = 1,225$  pairwise comparisons an 11.67-fold increase in computational cost. At current inference speeds (1-5 seconds per prompt with Llama-3.1-8B-instruct), processing 965 samples would require 328-1,640 hours on single-GPU infrastructure. Extending to 100 VLS yields  $C(100,2) = 4,950$  pairs a 47-fold increase over the current 105-pair configuration, rendering exhaustive pairwise evaluation computationally prohibitive for operational deployment.

Adaptive sampling strategies offer a viable approach to this scalability challenge. However, implementation must address a critical constraint: maintaining stratified input distribution while ensuring all VLS participate in sufficient comparisons to enable reliable ranking. Simple random subsampling risks creating disconnected comparison graphs where certain VLS receive insufficient evaluation, compromising rank estimation quality.

Proposed adaptive approaches could include:

- 1) Sequential Importance Sampling: Dynamically select pairwise comparisons based on current ranking uncertainty. Prioritize pairs involving VLS with high position variance or those near decision boundaries, reducing required comparisons while preserving ranking accuracy for critical items;
- 2) Hierarchical Decomposition: First-stage coarse clustering groups VLS by strategic category (sustainability, efficiency, innovation). Within-cluster ranking proceeds exhaustively (fewer pairs per cluster), followed by inter-cluster comparison using representative items. This approach reduces total comparisons from  $O(n^2)$  to approximately  $O(k(n/k)^2 + k^2)$ , where  $k$  represents cluster count, yielding substantial savings for large  $n$ ;
- 3) Graph Connectivity Guarantees: Implement minimum spanning tree algorithms to ensure each VL participates in at least  $\log(n)$  comparisons, preventing rank estimation

failure for under-sampled items. This constraint maintains statistical validity while enabling aggressive comparison reduction.

Future research should empirically validate these adaptive strategies against exhaustive pairwise evaluation using the v3.4 experimental configuration, benchmarking ranking correlation, position variance, and computational efficiency across VL set sizes ranging from 15 to 100.

## 6.2 Alternative Voting Aggregation Mechanisms

**Transitivity Violations and Cycle Analysis:** The pairwise comparison framework generates 105 comparisons for 15 VLs, creating substantial opportunity for intransitive preference cycles ( $A > B$ ,  $B > C$ , but  $C > A$ ). This study did not systematically quantify cycle frequency or assess their impact on final rankings - an acknowledged limitation. The simple vote-counting aggregation used here resolves cycles implicitly by accumulating wins without explicit cycle detection, potentially obscuring inconsistent model reasoning. Future research should: (1) quantify transitivity violation rates across experimental configurations, (2) evaluate whether cycle frequency correlates with per-VL variance artifacts, and (3) compare cycle-aware aggregation methods (Kemeny optimal ranking, Schulze method) against the baseline voting approach. High cycle rates would indicate fundamental instability in LLM pairwise judgments, while low rates would validate the aggregation methodology. This analysis represents a necessary extension for production deployment where ranking consistency guarantees are required.

The ranking determinism documented in section 4.B arises partly from pairwise voting mechanism - accumulating binary preferences across 105 comparisons that introduces systematic bias through consensus amplification. The following alternative mechanisms could mitigate this effect without requiring human intervention.

Alternative Aggregation Frameworks:

1) **Bradley-Terry Model:** Probabilistic framework estimating latent "strength" parameters  $\theta_i$  for each VL  $i$  by maximizing likelihood of observed pairwise outcomes. Under this model,  $P(i \text{ defeats } j) = \theta_i / (\theta_i + \theta_j)$ , enabling principled ranking even with inconsistent comparisons. Provides confidence intervals for rank positions and accounts for preference intensity. Computationally feasible via iterative maximum likelihood estimation using standard optimization libraries. The confidence intervals enable automated flagging of uncertain rank positions, replacing manual expert review with statistical thresholds for identifying unreliable output;

2) **Multi-Armed Bandit Approach:** Reformulates VL ranking as sequential decision problem where each comparison provides feedback signal. Upper Confidence Bound (UCB) or Thompson Sampling strategies prioritize informative comparisons - those maximizing expected information gain about uncertain rank positions. This adaptive strategy potentially reduces required comparisons from 105 to 40-60 while maintaining ranking quality, addressing scalability constraints identified in Section 5.A. Implementation requires defining reward function reflecting ranking accuracy and designing exploration-exploitation trade-off parameters. This self-correcting mechanism automatically concentrates evaluation effort on high-variance VLs, reducing position-specific bias without human identification of problematic items.

3) **Kemeny Optimal Ranking:** Seeks ranking minimizing total Kendall tau distance to observed pairwise preferences. This median ranking explicitly addresses transitivity violations by finding consensus ordering closest to all pairwise decisions. While NP-hard for general rankings, approximation algorithms (e.g., Borda count, Schulze method) provide tractable solutions for moderate-size problems ( $n \leq 50$ ). Particularly relevant for deployment scenarios requiring explainable aggregation logic. The median ranking computation smooths outlier preferences, mitigating high determinism that produced near-zero variance.

Empirical comparison of these alternatives should be benchmark: (1) correlation with v1 baseline (Kendall's  $\tau$ ), (2) internal consistency (position variance), (3) computational efficiency (required comparisons, runtime). Evaluation should use the v3.4 experimental configuration to isolate aggregation method effects from prompt engineering influences.

### 6.3 Model Family Comparison and Domain Adaptation Effects

While alternative voting addresses aggregation-induced bias, domain adaptation targets the second bias source: LLM architectural constraints affecting position or VL preferences. This research uses exclusively Llama-family models (3B, 8B parameters) due to inference cost constraints. The observed bias patterns: extreme variance shifts (Virtual Supplier Meetings (30.2→0.2); Sustainability Compliance (2.4→22.4)). This may reflect Llama-specific architectural characteristics or pre-training corpus composition rather than universal LLM behavior.

Cross-Architecture Validation: Future experiments should replicate the v3 series across alternative model families to assess generalizability:

1) GPT-4o/GPT-5: Proprietary models with broader pre-training corpora and RLHF alignment. Expected to exhibit different position biases due to distinct training paradigms. Higher inference costs limit exhaustive evaluation but enable targeted validation on high-variance VLs identified in Section 4.2;

2) DeepSeek-V3: Recent open model with competitive performance. Mixture-of-experts architecture may route different comparison types through specialized sub-networks, potentially reducing uniform bias patterns observed in dense Llama models;

3) Mixtral 8x7B: Another mixture-of-experts architecture with proven efficiency. Sparse activation patterns could mitigate softmax-like voting amplification effects documented in Section 4.2;

4) Gemma: Google's open model family. Cross-validation would reveal whether procurement ranking patterns generalize across fundamentally different training corpora and tokenization strategies.

Hypothesis: Models with broader pre-training corpora (GPT-5, Gemini) may exhibit less extreme position-specific artifacts than domain-constrained Llama variants. Ensemble methods combining multiple architectures could smooth individual biases through architectural diversity, analogous to model averaging in classical machine learning. However, ensemble approaches introduce computational overhead and require principled aggregation strategies to avoid compounding systematic errors.

Domain Adaptation Through Fine-Tuning: The position-specific biases documented in Section 5.2 suggest pre-trained LLMs lack nuanced procurement domain understanding. Targeted fine-tuning on procurement-specific ranking datasets (500-2,000 scenario-ranking pairs with expert annotations) could align model priors with domain expert distributions. As a simpler alternative, general-purpose fine-tuning on large procurement corpora or procurement-related conversational data could reduce domain misalignment without requiring task-specific annotation, though with potentially smaller bias reduction compared to ranking specific fine-tuning.

Expected Impact: Domain-adapted models should exhibit: (1) reduced position variance for unstable VLs (e.g., Sustainability Compliance), (2) improved correlation with expert baselines (weighted  $\tau > 0.45$ ), (3) better generalization to novel scenarios absent from training data, (4) explainable reasoning aligned with procurement frameworks rather than surface-level pattern matching.

## 6.4 Generalization Beyond Procurement Domain

Methodological insights likely transfer to analogous structured decision problems: healthcare treatment prioritization given patient profiles, financial investment strategy ranking under risk constraints, logistics route optimization with competing objectives. However, domain-specific validation remains essential before deployment.

Cross-domain research replicating core experiments (prompt format effects, context minimalism, tool integration) across 3-5 domains with distinct characteristics would identify universal LLM ranking principles versus context-dependent phenomena requiring domain-specific adaptation. Suggested validation domains: (1) clinical treatment selection (healthcare), (2) portfolio optimization (finance), (3) facility location planning (logistics), (4) R&D project prioritization (innovation management), (5) supplier selection (adjacent to procurement but distinct value criteria).

## 6.5 Deployment Considerations and Human-AI Collaboration

Despite identified limitations, LLM based ranking demonstrates operational viability for specific procurement use cases:

Appropriate Applications:

- 1) 84% position variance reduction across 965 scenarios (Section 4.D) proportionally reduces the scope of VL positions requiring fewer expert reviews;
- 2) Sensitivity analysis across scenario variations to identify high-impact strategic levers;
- 3) Training and calibration tool for junior procurement staff;
- 4) Hypothesis generation for strategic planning workshops.

Inappropriate Applications:

- 1) Fully automated decision-making without human oversight, given documented bias patterns;
- 2) Novel or edge-case scenarios outside training distribution (high variance VLs);
- 3) High-stakes decisions requiring auditability and regulatory compliance;
- 4) Real-time operational contexts requiring sub-second latency (current inference: 1-5 seconds per comparison).

Proposed Human-AI Collaboration Framework<sup>\*\*</sup>: Optimal deployment combines LLM strengths (consistency, scalability, context integration) with human expertise (edge case handling, ethical judgment, strategic intuition):

- 1) LLM Initial Ranking: System generates preliminary VL ordering with position-specific confidence scores derived from pairwise vote distributions;
- 2) Uncertainty Based Flagging: Algorithm identifies high-uncertainty positions (variance exceeding threshold) or VLs exhibiting extreme variance shifts (Section 4.B analysis);
- 3) Expert Validation: Procurement specialists review flagged positions, validating or adjusting based on contextual knowledge unavailable to pre-trained models;
- 4) Active Learning Loop: System learns from expert corrections, iteratively refining both model parameters (via fine-tuning) and prompt engineering strategies to reduce future uncertainty in similar scenarios.

This hybrid approach partially mitigates automation risks while capturing efficiency gains, positioning LLM based ranking as augmentation tool rather than replacement for human expertise in strategic procurement operations.

## 6.6 VL Naming Effects and Semantic Sensitivity

The per-VL variance patterns documented in Section 4.B raise a critical question: to what extent do VL names themselves, rather than underlying procurement logic-drive model ranking behavior

Future Research Directions for biased VLs:

- 1) Controlled Naming Experiments;
- 2) Synonym Sensitivity Analysis;
- 3) Domain Expert Alignment;
- 4) Multi-Category Evaluation: Extend analysis beyond PwP-related VLs to encompass 5-6 distinct procurement categories (commercial, technical, strategic, compliance, sustainability, innovation). The 13-20% artifact rate observed in this single-category study may represent category-specific model-terminology interactions. Multi-category portfolios with 50+ VLs exhibiting diverse semantic characteristics would reveal whether variance patterns generalize or dilute as category diversity increases, informing practical deployment scope.

## 6.7 Baseline Validity and Evaluation Framework Limitations

The use of a proprietary baseline (v1) as ground truth introduces reproducibility constraints that merit explicit acknowledgment. External researchers cannot independently verify v1 outputs or replicate exact correlation calculations. However, several factors mitigate this limitation:

First, the validation study (Section 3.C) establishes that v1 achieves substantial agreement with independent expert judgment (quadratic  $\kappa = 0.62$ ), providing empirical grounding for its use as benchmark. The observed LLM-baseline correlations (weighted  $\tau = 0.38-0.45$ ) should be interpreted relative to this expert-baseline agreement level: achieving  $\tau = 0.45$  against a baseline that itself has  $\kappa = 0.62$  expert agreement suggests LLM rankings approach the reliability ceiling imposed by inherent task ambiguity.

Second, the primary research questions effects of prompt engineering, categorical versus numeric representations, and strategic tool integration all evaluated through within-experiment comparisons that do not depend on v1 absolute accuracy. The documented 84% variance reduction (v3.3  $\rightarrow$  v3.4) derived independently of baseline validity.

Third, alternative evaluation approaches face their own limitations. Expert annotation of all  $965 \times 15 = 14,475$  VL positions would require prohibitive resources while introducing annotator fatigue effects. Synthetic ground truth lacks domain validity. Cross-model consensus baselines (averaging multiple LLM outputs) would confound evaluation with the phenomena under investigation.

Future research could strengthen validation through: (1) expanded expert panel assessment using stratified sampling, (2) comparative evaluation where experts rank both v1 and LLM outputs for the same scenarios, enabling direct preference elicitation, and (3) publication of input- output scenario-ranking pairs to enable partial reproducibility within proprietary constraints.

## 7.0 Conclusions

This study demonstrates that LLM based procurement VL ranking through pairwise comparison aggregation achieves operational viability for specific applications while revealing fundamental limitations that exclude fully automated deployment.

## 7.1 Key Empirical Findings

Correlation with proprietary baselines remained moderate across all configurations (weighted Kendall's  $\tau = 0.38-0.45$ ), indicating that LLM-generated rankings capture broad strategic patterns but diverge from domain expert judgments in systematic ways. The categorical value representation consistently outperformed numeric formats, particularly in smaller models, suggesting that LLM internal representations process qualitative descriptors more effectively than numerical scales for procurement decision contexts.

The most significant finding concerns internal consistency patterns. LLM based approaches exhibited dramatically higher determinism than the proprietary baseline, with position variance reductions ranging from 3-fold to 150-fold for specific VLs. This extreme stabilization arises from dual mechanisms: architectural biases in model representations and consensus amplification through the pairwise voting aggregation process.

Position-specific analysis revealed distributional shifts affecting specific VLs: Virtual Supplier Meetings variance decreased 150-fold (30.2→0.2), while Sustainability Compliance variance increased 9-fold (2.4→22.4). Importantly, these patterns admit multiple interpretations beyond model failure: (1) From procurement domain expertise, certain VLs represent near-universal best practices that experienced practitioners consistently prioritize regardless of scenario, such levers "almost always work" and may legitimately warrant stable high rankings; (2) The pairwise voting aggregation algorithm contributes to these effects, as no aggregation mechanism achieves perfect behavior across all items; (3) This study evaluated only PwP-related VLs - a single procurement category. Multi-category portfolios spanning 5-6 distinct VL types (commercial, technical, strategic, compliance, sustainability, innovation) would likely exhibit different variance distributions as semantic diversity dilutes category-specific patterns.

Strategic tool integration (Experiment v3.4) produced sixfold stability improvements compared to context-minimal configurations, validating the hypothesis that structured frameworks from established procurement methodologies enhance ranking consistency. However, this enhancement amplified deterministic behavior, raising concerns about reduced adaptability to novel scenarios requiring flexible expert judgment.

## 7.2 Methodological Contributions

The research establishes that removing VL numeric metadata from prompts (v3.3) reduces deterministic bias without catastrophic accuracy loss, achieving position stability (9.348) comparable to baseline levels while maintaining weighted correlation ( $\tau = 0.38$ ). This finding has practical implications: simpler prompts focusing on qualitative descriptors and contextual scenario information produce rankings less vulnerable to data-driven bias while preserving computational efficiency.

One more side effect is that some input data could be highly correlated to the input information, thus leading model to be confident in output and disregarding other important information.

The documented trade-off between consistency and flexibility reveals fundamental limitations of current LLM architectures for strategic decision support. High reproducibility desirable for systematic processes conflicts with the adaptive variability characteristic of expert decision-making across diverse procurement scenarios. This tension suggests that LLM based ranking functions optimally as augmentation rather than replacement for human expertise.

## 7.3 Practical Viability Assessment

Proposed solution demonstrates clear utility for initial screening, sensitivity analysis achieving 84% reduction in ranking position variance across 965 scenario configurations (position stability: 9.348 → 1.465, Section 4.D), which proportionally reduces the number of VL positions requiring substantive expert assessment based on consistency improvements.

However, documented bias patterns and deterministic behavior establish critical boundaries: deployment requires human oversight for high-stakes decisions, edge-case scenarios, and contexts demanding final human review.

The study contributes a validated human-AI collaboration framework that leverages computational consistency alongside domain expertise. This hybrid approach combining LLM generated preliminary rankings with uncertainty based expert validation involvement mitigates automation risks while capturing efficiency gains.

Additionally, this study finds that even simple metrics like stability and case study distribution analysis are critical before deployment due to potential critical bias in results.

## 7.4 Broader Implications

This research yields insights extending beyond procurement to LLM based decision support systems generally. The dual-mechanism bias pattern (architectural constraints combined with aggregation amplification) likely manifests in any pairwise LLM comparison frameworks, independent of application domain. The categorical-versus-numeric representation effect demonstrates that LLM internal representations encode qualitative descriptors differently from numerical values, with implications for any multi-criteria evaluation application requiring numerical reasoning.

The context minimalism principle emerging from experiment v3.3 contradicts intuitions favoring comprehensive information provision. Removing detailed metadata reduced deterministic bias while preserving ranking quality, suggesting that excess context may trigger pattern-matching rather than substantive reasoning. This finding applies to analogous domains.

Model calibration with respect to bias and stability should constitute a mandatory component of any LLM-related deployment pipeline. Beyond standard accuracy metrics, practitioners should incorporate domain-specific calibration objectives reflecting task requirements. The experimental methodology combining accuracy metrics (Kendall's  $\tau$ ) with position variance analysis provides a replicable framework for detecting bias patterns invisible to accuracy-only evaluation, establishing recommended practice for LLM decision support validation across domains.

## References

- [1] Singh, P. K., & Chan, S. W. (2022). The Impact of Electronic Procurement Adoption on Green Procurement towards Sustainable Supply Chain Performance-Evidence from Malaysian ISO Organizations. *Journal of Open Innovation: Technology, Market, and Complexity*, 8(2). <https://doi.org/10.3390/joitmc8020061>
- [2] Gurgun, A. P., Kunkcu, H., Koc, K., Arditi, D., & Atabay, S. (2024). Challenges in the Integration of E-Procurement Procedures into Construction Supply Chains. *Buildings*, 14(3), 605. <https://doi.org/10.3390/buildings14030605>
- [3] Cui, R., Li, M., & Zhang, S. (2020). AI and Procurement. *Manufacturing & Service Operations Management*. <https://doi.org/10.2139/ssrn.3570967>
- [4] Moenks, N., Penava, P., & Buettner, R. (2025). A Systematic Literature Review of Large Language Model Applications in Industry. *IEEE Access*, 13. <https://doi.org/10.1109/ACCESS.2025.3608650>
- [5] Chan, A. P. C., & Owusu, E. K. (2022). Evolution of Electronic Procurement: Contemporary Review of Adoption and Implementation Strategies. *Buildings*, 12(2), 198. <https://doi.org/10.3390/buildings12020198>
- [6] Afolabi, A., Ibem, E., Aduwo, E., Tunji-Olayeni, P., & Oluwunmi, O. (2019). Critical Success Factors (CSFs) for e-Procurement Adoption in the Nigerian Construction Industry. *Buildings*, 9(2), 47. <https://doi.org/10.3390/buildings9020047>

- [7] Ye, Y., Zhang, Z., Ma, T., Wang, Z., Li, Y., Hou, S., Sun, W., Shi, K., Ma, Y., Song, W., Abbasi, A., Cheng, Y., Cleland-Huang, J., Corcelli, S., Culligan, P., Goulding, R., Hu, M., Hua, T., Lalor, J., Liu, F., Luo, T., Maginn, E., Moniz, N., Rohr, J., Savoie, B., Slate, D., Stapleford, T., Webber, M., Wiest, O., Zhang, J., & Chawla, N. V. (2025). LLMs4All: A review on large language models for research and applications in academic disciplines. arXiv. <https://doi.org/10.48550/arXiv.2509.19580>
- [8] Pesch, P. J., Hofmann, H. C. H., & Pflücke, F. (2025). Potentials and challenges of large language models (LLMs) in the context of administrative decision-making. *European Journal of Risk Regulation*, 16, 76–95. <https://doi.org/10.1017/err.2024.99>
- [9] Aboelazm, K. S., & Dganni, K. M. (2025). Public procurement contracts futurity: Using of artificial intelligence in a tender process. *Corporate Law & Governance Review*, 7(1), 60-72. <https://doi.org/10.22495/clgrv7i1p6>
- [10] Andhov, M., Darnall, N., & Andhov, A. (2025). Leveraging AI for sustainable public procurement: opportunities and challenges. *Frontiers in Sustainability*, 6, 1603214. <https://doi.org/10.3389/frsus.2025.1603214>
- [11] Anghel, C., Anghel, A. A., Pecheanu, E., Cocu, A., Istrate, A., & Andrei, C. A. (2025). Diagnosing Bias and Instability in LLM Evaluation: A Scalable Pairwise Meta-Evaluator. *Information*, 16(8), 652. <https://doi.org/10.3390/info16080652>
- [12] Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, S., Zhang, K., Wang, Y., Gao, W., Ni, L., & Guo, J. (2025). A Survey on LLM-as-a-Judge. arXiv. <https://doi.org/10.48550/arXiv.2411.15594>
- [13] Abdelkarim, S., Lu, D., Flores, D., Jaeggi, S., & Baldi, P. (2025). Evaluating the Intelligence of large language models: A comparative study using verbal and visual IQ tests. *Computers in Human Behavior: Artificial Humans*, 5, 100170. <https://doi.org/10.1016/j.chbah.2025.100170>
- [14] Baeza-Yates, R., & Ribeiro-Neto, B. (2011). *Modern Information Retrieval: The Concepts and Technology behind Search* (2nd ed.). Addison-Wesley.
- [15] Ricci, F., Rokach, L., & Shapira, B. (2011). *Introduction to Recommender Systems Handbook*. Springer. [https://doi.org/10.1007/978-0-387-85820-3\\_1](https://doi.org/10.1007/978-0-387-85820-3_1)
- [16] Yilmaz, E., & Aslam, J. A. (2008). Estimating average precision when judgments are incomplete. *Knowledge and Information Systems*, 16, 173–211. <https://doi.org/10.1007/s10115-007-0101-7>
- [17] Webber, W., Moffat, A., & Zobel, J. (2010). A similarity measure for indefinite rankings. *ACM Transactions on Information Systems*, 28(4). <https://doi.org/10.1145/1852102.1852106>