Categorization Architecture with Predictive Reasoning and Alignment for UNSPSC

Erli Wang, Dakshi Kapugama Geeganage, and Opec Kemp

Queensland Government Procurement, Department of Housing and Public Works,
Queensland Government, Australia
{Erli.Wang, Dakshi.KapugamaGeeganage, Opec.Kemp}@hpw.qld.gov.au

Abstract. The United Nations Standard Products and Services Code (UNSPSC) is a four-tier hierarchical taxonomy comprising segments, families, classes and commodities, designed to facilitate globally consistent categorization of goods and services. The Queensland Government quarterly processes over seven million new procurement transactions, all of which, in line with recommendations from the Queensland Audit Office, are mapped to the UNSPSC hierarchy to support timely decision-making and improve operational efficiency. In this paper, we introduce the Categorization Architecture with Predictive Reasoning and Alignment (CAPRA), a framework for rigorously evaluating real-world labeling pipelines. CAPRA systematically compares multiple candidate configurations across four key dimensions: accuracy, model size, inference speed and API cost. To maintain data quality without overburdening human experts, who cannot feasibly validate millions of transactions each quarter, we embed an AI-driven feedback mechanism and a retrieval-based review process that selectively flags only the lowest-confidence predictions for expert adjudication. This "selective confidence-based review" closes the data-service loop by focusing domain expertise where it matters most. Finally, we demonstrate CAPRA's effectiveness through extensive experiments on the Queensland Government expenditure. Our results show that CAPRA outperforms a fine-tuned LLM by 19% in prediction accuracy while requiring only 10% of its inference time, enabling seamless deployment in a production environment.

Keywords: Government Procurement · UNSPSC · Semantic Similarity · LLM · Machine Learning Architecture · Multiclass Prediction.

1 Introduction

Government procurement constitutes a substantial portion of public expenditure and plays a pivotal role in delivering services and infrastructure to communities. For instance, the Queensland Government has annual procurement outlays exceeding 25 billion Australian dollars. Despite an agency-led model that sets strategic priorities through six Procurement Categories — GGS (General Goods and Services), TIS (Transport Infrastructure Services), ICT (Information and Communication Technology), BCM (Building Construction and Maintenance), SOC (Social Services) and MED (Medical) —, each individual agency retains autonomy over its own data quality and purchasing decisions. This decentralised model, though flexible, often leads to inconsistent classification, data fragmentation, and challenges in cross-agency spend analysis.

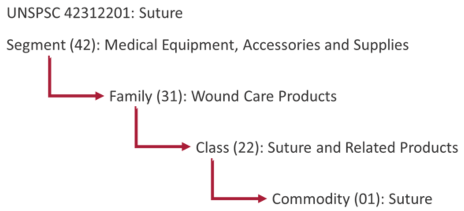
UNSPSC 42312201: Suture
Segment (42): Medical Equipment, Accessories and Supplies
Family (31): Wound Care Products
Class (22): Suture and Related Products
Commodity (01): Suture

Fig. 1: An Example in UNSPSC

A robust, unified taxonomy is therefore essential to ensure data integrity and enable meaningful comparisons across disparate procurement systems. The United Nations Standard Products and Services Code (UNSPSC) taxonomy [10] is constructed as a tree structure with four levels: segment, family, class, and commodity (see in Figure 1). At each level, categories become progressively more granular. The UNSPSC harmonise product and service records across agencies. The Queensland Audit Office (QAO) has formally recommended the implementation of UNSPSC to enhance transparency, eliminate duplications, and support more effective oversight of procurement activities.

These advantages set the stage for leveraging AI to automate UNSPSC coding at scale, thereby closing the loop between data consistency and value-driven procurement. Various formulation and its corresponding approaches have been explored to automate UNSPSC categorization. In [9], multiple zero-shot prompting strategies have been test on a dataset of 50 000 diverse inventory items. It shows that LLMs can automate UNSPSC classification with 40.3 % class-level and 54.6 % segment-level accuracy—and when provided context on a focused 27,000-item segment, performance rises to 72.6 % class-level and 90.3 % segment-level. A variety of machine learning have been tested on UNSPSC categorization [1], including Logistic Regression, Support Vector Machine, Random Forest, XGB, MLP, and LSTM. Most of models are able to achieve 85+% accuracy, however, the variety of labels in the experiment are small, ranging from 8 classes to 11 classes. Similar to UNSPSC, [2] show that the state-of-the-art LLMs (Claude 3.5 Sonnet and GPT-4) sustain robust classification of agricultural products into 8-digit Harmonized System codes, achieving accuracy rates ranging between 60% and 90% depending on HS aggregation levels.

Considering aforementioned recommendations, benefits, and limitations of small-scale classification for products and services, we designed Categorization Architecture with Predictive Reasoning and Alignment (CAPRA) for UNSPSC classification. The contributions of this paper are twofold. First, we provide a comprehensive large-scale comparison, ranging from zero-shot and fine-tuned LLMs to semantic-embedding with tree-based classifiers, under a unified experimental framework. Second, we introduce and evaluate an AI-feedback–driven validation phase to flag only the lowest-confidence predictions for human review. This selective human-in-the-loop design maintains high data integrity while dramatically reducing expert workload.

2 Data Preparation for UNSPSC Categorization

Inspired by California's publicly available purchase order data [3], we collected procurement transactions from the Queensland Government using SAP (Systems, Applications, and Products) systems, corporate card platforms, and agency-specific sources. Each quarter, a robust ETL (Extract, Transform, Load) process ingests over 18GB of data from more than 4,000 files across 22 agencies operating on 10 different ERP (Enterprise Resource Planning) systems. This includes over 25 million transaction records linked to more than 100,000 general ledger codes, which are distilled into approximately 1.6 million unique entries for UNSPSC classification.

After consolidation, we apply a multistep data preparation pipeline that includes preprocessing, feature extraction, and transformation to convert raw procurement data into structured and meaningful inputs for machine learning.

2.1 Feature Engineering

Not all available features contribute equally to classification performance. To improve model accuracy and robustness, we applied feature selection, feature combination, and noise reduction using exploratory analysis and domain knowledge. Our dataset included fields such as supplier name, supplier type (public/private), and agency-specific metadata. However, many were noisy, sparse, or weakly correlated with the target UNSPSC labels.

Feature Selection: We performed both statistical and domain-informed selection, assessing correlation with target labels and semantic value for embedding-based models. Three features were retained:

- **GLDescription**: General ledger description of the procurement item
- **LineText**: Raw transaction-level line item text
- **ABRIndustry**: Supplier's industry from Australian Business Register (ABR)

These features collectively provide product and supplier context, critical for accurate UNSPSC prediction. Table 1 shows sample records.

Table 1: Sample Procurement Line Items with UNSPSC Titles

GLDescription	LineText	ABRIndustry	UNSPSC Title
Repairs & maintenance Buildings	Building(Mary) INV0000	House Construction	Building maintenance service
Security	P/O XXX INV0000 CASHGUARD	Investigation and Security Services	Security guard services

Feature Combination: To enrich the semantic input for the embedding model, we combined the selected fields *GLDescription*, *LineText*, and *ABRIndustry* into a single concatenated string for each record. This combination allowed the model to learn from complementary contextual cues across features.

Noise Reduction and Pruning: We removed features that were sparse, inconsistent, or weakly informative, such as *SupplierName* and certain agency metadata. Excluding these fields helped improve generalisability and reduced distractions for the model.

By focusing on a clean and informative feature set, we improved the quality of embeddings and enhanced downstream prediction accuracy.

2.2 Mitigating Class Imbalance through Data Augmentation and Stratified Sampling

The initial labelled dataset, primarily sourced from the Department of Transport and Main Roads (TMR), was heavily skewed toward transport and infrastructure services, resulting in significant class imbalance across UNSPSC categories. To improve representation of other major procurement categories such as GGS, BCM, and ICT, we curated additional data from Queensland Health (medical procurement), QBuild (construction and maintenance), and a gold standard dataset manually labelled by domain experts based on initial model predictions.

To address class imbalance and ensure broad coverage of procurement categories, we constructed a more balanced dataset using stratified sampling. This was applied for two key reasons: first, to improve the overall distribution of UNSPSC labels across training and testing sets, and second, to promote diversity within each major procurement category by ensuring adequate representation of unique UNSPSC titles. We selected stratified samples from key procurement domains, as shown in Table 2, which summarises the number of labelled instances per category in the final training dataset.

Table 2: Label Distribution in the Final Training Dataset

Category	**Number of Labels**
General Goods and Services (GGS)	722
Medical (MED)	571
Transport and Infrastructure Services (TIS)	347
Building and Maintenance (BCM)	314
Information and Communication Technology (ICT)	180

3 CAPRA: Categorization Architecture with Predictive Reasoning and Alignment

As illustrated in Figure 2, incoming procurement records are first encoded into dense semantic embeddings via a pretrained language model, capturing the nuanced meaning of each description. These embeddings then drive a supervised classifier, trained and validated on historical transactions, to predict the corresponding UNSPSC code. Immediately after prediction, we compute two complementary confidence measures: a Retrieval-Augmented Value (RAV) that

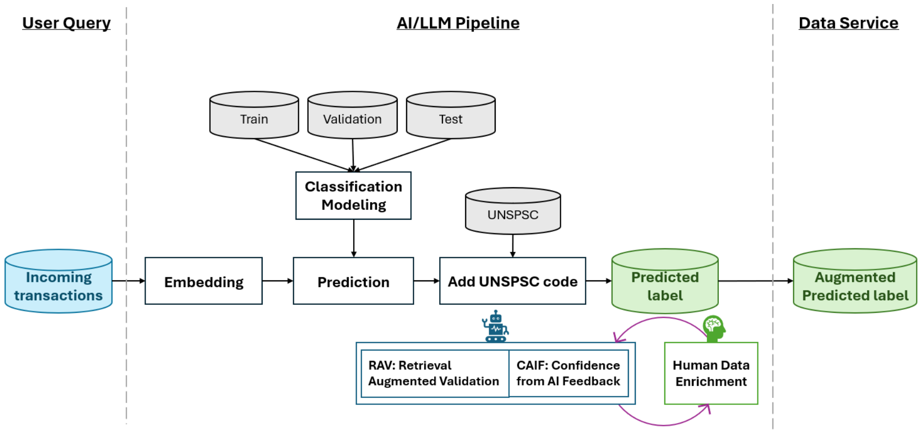

Fig. 2: The inputs, outputs, and main components of the proposed CAPRA.

flags transactions rarely seen in our corpus, and a GPT-based plausibility score (CAIF) that assesses the semantic coherence of each description–code pairing.

Transactions with low RAV or low CAIF are the only ones routed to a streamlined review interface, where domain specialists swiftly confirm or correct the code. Their feedback is reincorporated into the training set, continuously refining the classifier and driving ever-improving accuracy. This closed-loop design combines automated embedding-based categorization for routine volumes with targeted expert intervention for edge cases, delivering scalable, high-quality UNSPSC labeling across millions of records.

In the following subsections, we describe details of Supervised Classification, Retrieval-Augmented Validation (RAV), Confidence from AI Feedback (CAIF).

3.1 Classification module

Let the training dataset be

$$\mathcal{D}^{\mathrm{tr}} = \{(x_i, y_i)\}_{i=1}^{n}, \quad x_i \in \mathbb{R}^D,\ y_i \in \mathcal{Y} = \{1, \dots, C\},$$

where each input x_i is a D-dimensional feature vector and y_i its class label. In our collected dataset, not all entries have the full 4-layer structure from commodity to segment, therefore, we flat out labels to all four levels. In other words, y_i is made to be the most granular label we can find. In term of x_i, it is represented as a list of open-text, such as ['Repairs & maintenance – Buildings', '123456781 Bradfield Building I:INV-0001', and 'House Construction'].

An unseen test dataset $\mathcal{D}^{\mathrm{te}} = \{(x_j, y_j)\}_{j=1}^{m}$ shares the same input and label spaces but is not available during training. Our objective is to learn a classifier $f_\phi : \mathbb{R}^D \to \mathcal{Y}$, parameterized by ϕ, that minimizes the expected misclassification error on the test distribution: $\min_\phi \mathbb{E}_{(x,y)\sim\mathbb{P}^{\mathrm{te}}}\big[\mathbf{1}\{f_\phi(x) \neq y\}\big]$.

LLM-Based Zero-Shot Classification. The zero-shot approach leverages a pretrained LLM to assign UNSPSC codes directly without task-specific training.

- Instruction-Driven Prompt. For example:

```
f"""Classify this procurement item into the most appropriate UNSPSC
title. Respond with only the most relevant UNSPSC title. Details:
- GL Description: {row['GLDesc']}
- Line Text: {row['LineText']}
- ABR Industry: {row['ABRIndustry']}
"""
```

- Label-List Conditioning. For example:

```
f"""Classify the UNSPSC title. Respond only with the correct UNSPSC
Final Title from the following
options: {unspsc_titles_str}. Details:
- GL Description: {row['GLDesc']}
- Line Text: {row['LineText']}
- ABR Industry: {row['ABRIndustry']}
```

This approach requires no labeled data or fine-tuning, can adapt to new codes on the fly, but suffers from high latency, sensitivity to prompt phrasing, and occasional hallucinations. Considering the high volume of transactions, we do not involve advanced prompting techniques, such as Chain-of-Thought, or Self-Consistency voting for the time and cost constraints.

LLM-Based Fine-Tuned Classification. The fine-tuned approach starts from a pretrained LLM that is further optimized on a domain-specific UNSPSC-annotated dataset via supervised learning based on given corpus of ⟨description, code⟩ pairs.

At inference, the fine-tuned model requires no candidate enumeration: we feed it the cleaned transaction text, and it emits the most likely UNSPSC code in a single forward pass. Fine-tuning normally yields higher in-domain accuracy and consistency than zero-shot but incurs additional training cost.

- The prompting at inference:

```
f"Predict the UNSPSC category, based on:
- Line Text: {row['LineText']}
- GL Description: {row['GLDesc']}
- Supplier Industry: {row['ABRIndustry']}"
```

Tabular Classification with Pretrained Embeddings. This hybrid pipeline first encodes each text description into a fixed-length vector embedding using a pretrained encoder and then applies a lightweight tabular classifier [7] to predict the UNSPSC code. The main worklfow is:

- Embed x into $z = f_\theta(x)$ using a model like `gte-Qwen2-1.5B-instruct` (e.g., "Elevator maintenance inspection" → $z \in \mathbb{R}^D$).
- Train a multiclass classifier on $\{(z_i, y_i)\}$, where y_i are the ground-truth codes.
- At inference, compute z the embedding of a new transaction and apply the classifier to obtain a label that the model has a high confidence.

Balancing accuracy, latency, size, and cost, Section 4 evaluates these methods on procurement data to identify the optimal scalable UNSPSC categorization.

3.2 CAIF: Confidence from AI Feedback

Human expert validation is the gold standard for model quality but is not likely to scale to the 7 million procurement records processed quarterly in Queensland. Moreover, procurement classification is more complex than simple binary tasks: overlapping categories and nuanced descriptions demand consistency with a standardized UNSPSC convention, not just semantic "correctness." Therefore, our goal is maximizing agreement with domain experts' annotations.

Recent studies show modern LLMs can match human evaluators' judgments [4, 6], so we introduce an LLM-driven feedback mechanism that uses the model's relevance score as a confidence signal and generates human-readable rationales for explainability. Below is the prompt template to assess description–code correspondence:

```
prompt = f"""
You are an expert in UNSPSC procurement classification. The model predicted
the category: "{cat_pred}". Given the following transaction details:
- GLDesc: {gl_desc}
- abrindustry: {abrindustry}
- LineText: {line_text}

Please evaluate how confident you are that this prediction is correct
on a scale from 0 to 10. Return a JSON object with two fields:
"confidence": <float between 0 and 10>,
"comment": <concise rationale for your score>.
"""
```

3.3 Retrieval-Augmented Validation (RAV)

To assess whether a predicted UNSPSC label is well supported by historical data, we introduce Retrieval-Augmented Validation (RAV). First, we embed every labeled transaction in our corpus, and index these vectors with FAISS [5] for fast approximate nearest-neighbor search. At test time, each new transaction is likewise embedded and we retrieve its top-k nearest neighbors by cosine similarity. We then compute the RAV@k score as the fraction of those k neighbors whose true labels match the model's prediction

Suppose the training set is $\mathcal{D}^{\text{tr}} = \{(x_i, y_i)\}_{i=1}^{N}$. For a pair of prediction (x', y'), the RAV@k score is defined as,

$$\text{RAV@}k = \sum_{i=1}^{N} \mathbb{I}_{\{\texttt{Similarity}(x',\, x_i) \geq \gamma\}} \cdot \texttt{Matches}\left(y', y_i\right)$$

where γ is the least similarity score in top-k, the $\texttt{Similarity}(x', x_i)$ is obtained from FAISS, and $\texttt{Matches}()$ simply exact match.

The RAV@k score measures the frequency of test instances for which the predicted label is sufficiently supported among the top-k retrieved labels. A high RAV@k indicates that the predicted code is corroborated by similar historical cases, adding an explainable, context-aware layer to standard accuracy metrics.

4 System Verification

Our final dataset, compiled from multiple procurement sources and refined through expert validation and stratified sampling, contained a total of 258,278 labelled records. We applied an 80:20 stratified split based on the final UNSPSC titles to preserve class distribution in both subsets. This resulted in 207,635 instances for training and 50,643 for testing, ensuring representative coverage of procurement categories.

4.1 Question 1: Does the classification model aligned with annotator in an efficient way?

The goal of our experiments is to identify the most suitable technical methodology to automate the UNSPSC categorization at scale. To this end, we evaluate it across four key dimensions: classification accuracy, model size, inference speed, and cost (Table 3).

Table 3: Performance Comparison. The cost is based on [8].

Approach	**Accuracy**		**Time**			**API**	**Size**
	Train	Test	Emb.	Model	Inference		
`GPT-4 Zero-shot`	–	5%	–	–	7 hrs	$70	–
`GPT-4 Zero-shot with label list`	–	31%	–	–	9 hrs	$3689	–
`GPT-3.5-turbo fine-tuning`	–	73%	–	4 hrs	11 hrs	$300	–
`text-embedding-3-small + RF`	98%	90%	0.7 hr	2 hrs	1.0 hrs	$0.36	105 G
`text-embedding-3-small + LightGBM`	99%	92%	0.7 hr	2 hrs	1.1 hrs	$0.36	0.87 G
`gte-Qwen2-1.5B-instruct + LightGBM`	99%	92%	1 hr	2 hrs	1.1 hrs	–	0.84 G

Embedding-based classification pipelines consistently outperform LLM-only approaches in both accuracy and operational efficiency: on our UNSPSC procurement dataset, `text-embedding-3-small`+`LightGBM` achieved 92 % test accuracy—19 pp higher than the best fine-tuned `GPT-3.5-turbo` (73 %)—training in 4 hrs, inferring in 1.1 hrs, and costing just $0.36 in API fees; in contrast, `GPT-4` zero-shot languished at 5 – 31 % accuracy after 7 – 9 hrs of inference and up to $3,700 in fees. Embedding models project descriptions into a vector space where semantic closeness mirrors UNSPSC categories, and `LightGBM` carves decision boundaries across hundreds of classes, whereas generative LLMs falter—zero-shot `GPT-4` yields near-random labels among 150 000 possibilities (only 1 % relevant), and even fine-tuning `GPT-3.5-turbo` on that 1 % subset (boosting accuracy from 31 % to 73 %) still trails far behind at much higher cost.

In imbalanced multilabel problems like this UNSPSC dataset, Accuracy alone may report overly high numbers just by predicting the majority labels. Table 4 compares three classifiers on macro-averaged metrics for UNSPSC categorization. The `text-embedding-3-small + RF` (`RF` stands for `RandomForest`) baseline achieves 90% accuracy with a macro precision of 0.77, macro recall of 0.64, and macro F1 of 0.68. Replacing `RandomForest` with `LightGBM` (while keeping the same embeddings) increases accuracy to 92% and boosts macro recall by 4 points (from 0.64 to 0.68), yielding a modest macro F1 gain to 0.69 despite a slight dip in precision (0.74 vs. 0.77). Finally,

Table 4: Classification performance comparison across different models

Approach	**Acc.**	**Macro Precision**	**Macro Recall**	**Macro F1**
`text-embedding-3-small + RandomForest`	0.90	**0.77**	0.64	0.68
`text-embedding-3-small + LightGBM`	0.92	0.74	0.68	0.69
`gte-Qwen2-1.5B-instruct + LightGBM`	0.92	**0.77**	**0.72**	**0.73**

swapping in the larger `gte-Qwen2-1.5B-instruct` embeddings with `LightGBM` achieves the 92% accuracy, restores macro precision to 0.77, and further elevates macro recall to 0.72—driving the highest macro F1 of 0.73. Overall, these results demonstrate that `LightGBM` consistently outperforms `RandomForest` on the same embeddings, and that richer embeddings yield the most balanced improvements in precision and recall.

4.2 Question 2: How to Derive Transaction-Level Confidence to Avoid Overburdening Domain Experts?

Although our models achieved over 90% accuracy overall, such aggregate metrics mask per-transaction uncertainty and do not guarantee reliability at the individual record level. In practice, we submit model predictions to domain experts for validation—but human review of multi-million transactions in a few days is neither scalable nor guaranteed error-free. Thus, maintaining high data quality without overburdening experts requires a selective, confidence-aware adjudication process.

To address this, we propose an AI-driven feedback and retrieval-based review framework in Table 5 to identify a small subset of lowest-confidence predictions where expert efforts matter the most.

Table 5: Confidence matrix combining RAV and LLM judgments.

	CAIF Judge Low	**CAIF Judge High**
RAV High	Familiar but Questionable (High RAV, Low CAIF)	Safe Zone (High RAV, High CAIF)
RAV Low	Rare & Implausible (Low RAV, Low CAIF)	Novel but Plausible (Low RAV, High CAIF)

Under this framework, two complementary scores narrow down to a handful transactions for category experts to review.

- **Retrieval-Augmented Value (RAV)**: For each transaction, RAV counts historical records with matching descriptions and predictions. High RAV suggests stability from ample precedent; low RAV flags are cases for review. (see Table 6).
- **Confidence from AI Feedback (CAIF)**: We prompt a LLM to assign a confidence rating (0–10) reflecting the plausibility of each description–prediction pair based on its general knowledge. High CAIF scores imply strong semantic correlation, while low scores mark potential misclassifications (See Table 7).

Table 6: RAV Score and Feedback on Transaction-Level UNSPSC Predictions

GLDesc	LineText	ABRIndustry	Prediction	RAV Score	RAV Justification Summary
Managed Items - Office Equipment	FUJIFILM #I002 APEOS AC2560	Other Machinery and Equipment Wholesaling n.e.c.	Photocopiers	10	High confidence: Predicted label is well supported. Found 10/10 matches for predicted label 'Photocopiers'. Top labels retrieved: Photocopiers (10)
Miscellaneous Plant	RGM 05.09.23 Inv. 001#1234	Other Automotive Repair and Maintenance	Flatbed truck	7	High confidence: Predicted label is well supported. Found 7/10 matches for predicted label 'Flatbed truck'. Top labels retrieved: Flatbed truck (7), Earthmoving machinery parts and accessories (2), Heavy equipment maintenance and repair service (1)
Subscriptions	prepayment FY2024.25 Subscriptions	Magazine and Other Periodical Publishing	National postal delivery services	3	Partial match: Some support for predicted label. Found 3/10 matches for predicted label 'National postal delivery services'. Top labels retrieved: National postal delivery services (3), Newspapers (3), Personnel relocation (3)

Table 7: CAIF Score and Feedback on Transaction-Level UNSPSC Predictions

GLDesc	LineText	ABRIndustry	Prediction	GPT Confidence	GPT COmments
Stores & Stationery	BCC OFFICE CHOICE STORES & STATIONERY	Stationery Goods Retailing	Office supply	10	All transaction details explicitly reference stores and stationery, and the industry is stationery goods retailing. "Office supply" is the correct UNSPSC category
INVENTORY–DRUGS PHARMACY I–PHARMACY	6745081	Chemist or Pharmacy Operations	Agents affecting water and electrolytes	6	The GLDesc and industry indicate a pharmacy or drug inventory context, which could include agents affecting water and electrolytes, but the information is too generic to confirm this specific drug class. The LineText '6745081' does not provide further clarification. The prediction is plausible but not certain.
COS Sub Contract	(08/08/2022) REPLACE FAULTY LIGHT SWITCH	Electrical Services	Electrical inspection service	2	The transaction describes the replacement of a faulty light switch, which is a repair or maintenance activity, not an inspection. The correct UNSPSC category would likely be 'Electrical system services' or 'Electrical maintenance or repair service', not 'Electrical inspection service'.

To address prediction reliability, we introduce a quadrant framework partitioning scenarios by model confidence and historical support. To have a better visualization of how the confidence quadrant works, we sample 100 transactions from FY23/24. Figure 3 shows a scatter plot: 40% in the safe zone, 30% needing expert input, and 30% in regions of intrinsic difficulty. By combining RAV and CAIF, our framework dynamically prioritizes human review for transactions that are both uncommon and lack semantic support, ensuring data integrity without overwhelming domain experts.

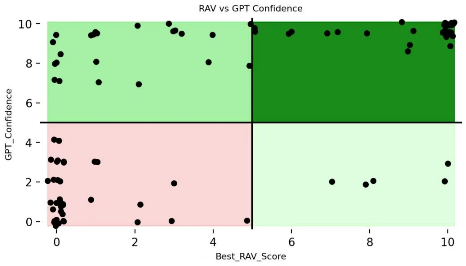

Fig. 3: Prediction Quality Distribution

4.3 Question 3: How can UNSPSC categorization enable more granular analysis of government expenditure?

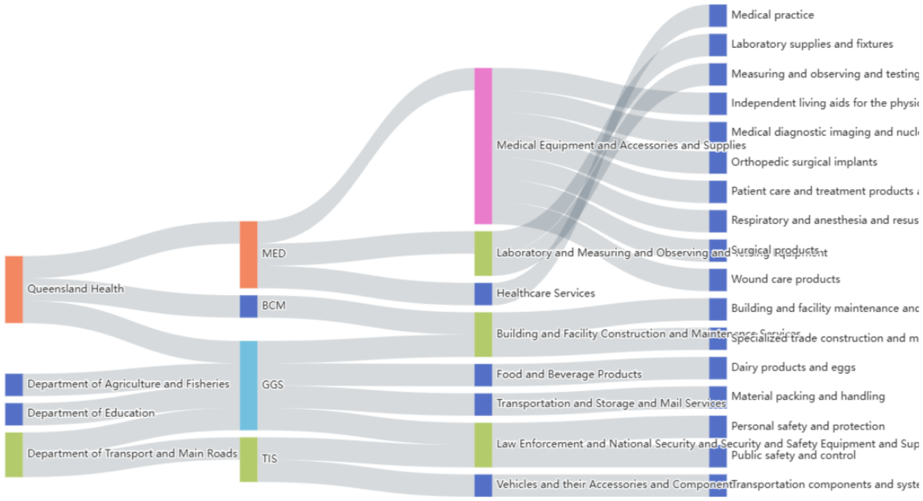

Fig. 4: Government expenditure tracking to class level with (RAV=5, CAIF=5)

Figure 4 depicts a Sankey diagram of government procurement flows, showing how expenditures move from issuing agencies (left) into high-level UNSPSC segments (center), and finally into detailed UNSPSC classes (right). The width of each flow is scaled to the total spend, immediately highlighting major channels—such as Queensland Health's substantial investment in "Medical Equipment & Accessories and Supplies," which further disaggregates into categories like "Laboratory Supplies and Fixtures" and "Orthopedic Surgical Implants." Similarly, the Department of Transport and Main Roads directs the lion's share of its budget into "Vehicles and their Accessories and Components," underscoring its transportation-centric purchasing profile.

This granular visualization is only achievable with a high-fidelity UNSPSC classification model. Precise, transaction-level labeling ensures that each procurement line item is correctly mapped to its appropriate commodity segment and class. Without such accuracy, misclassifications would propagate through the Sankey flows, obscuring true spending patterns and potentially leading to misguided policy interventions or budgetary allocations.

By revealing the actual distribution of public funds, this diagram delivers substantial public value. It enhances transparency and accountability in fiscal stewardship, empowers procurement officers to pinpoint opportunities for cost savings, and facilitates strategic supplier consolidation. Moreover, policymakers can leverage these insights to monitor compliance with budget guidelines, prioritize high-impact investments, and bolster trust in government operations—thereby driving more efficient resource allocation and reinforcing the integrity of public procurement processes.

5 Summary

We evaluate six categorization approaches—ranging from zero-shot and fine-tuned LLMs (`GPT-3.5-turbo`, `GPT-4`) to embedding-based pipelines (`text-embedding-3-small` and `gte-Qwen2` with `RandomForest` and `LightGBM`)—on a large Australian procurement dataset. Embedding + LightGBM delivers the best trade-off (92 % accuracy, 1.1 hr inference, <1 GB model, no API cost), vastly outperforming LLM methods (5 %-73 %). To ensure per-transaction reliability without overwhelming experts, we introduce a human-in-the-loop review that flags low-confidence cases using RAV and AI-derived confidence scores. Finally, we demonstrate how high-quality classification enables actionable Sankey visualizations of spend flows, enhancing transparency, driving cost-savings, and supporting data-driven procurement policy. We hope that this new advancement will further advance the practicality of AI adoption in public sector and allow more widespread applications of this robust approach to decision-making at scale.

Acknowledgments. We thank Andrew Bennett and Peter Lacey for agreeing to support this work. We are also grateful to Brent Phillips and Dr Daniel Philip for their countless constructive discussions and strategic insights. Finally, we thank Belinda Pfeffer, Susan Greaves, Sherling Ong, and Peita Cowley for their efforts in analysing, collecting, and developing the Queensland Government Procurement Data Warehouse.

References

1. Abdullahi, B., Ibrahim, Y.M., Ibrahim, A.D., Bala, K., Ibrahim, Y., Yamusa, M.A.: Development of machine learning models for classification of tenders based on UNSPSC standard procurement taxonomy. International Journal of Procurement Management **19**(4), 445–472 (2024)
2. Marra de Artiñano, I., Riottini Depetris, F., Volpe Martincus, C.: Automatic product classification in international trade: Machine learning and large language models. Review of International Economics (2025)
3. DGS: Purchase order data, https://data.ca.gov/dataset/purchase-order-data
4. Ding, B., Qin, C., Liu, L., Chia, Y.K., Joty, S., Li, B., Bing, L.: Is GPT-3 a good data annotator? arXiv preprint arXiv:2212.10450 (2022)
5. Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.E., Lomeli, M., Hosseini, L., Jégou, H.: The faiss library. arXiv preprint arXiv:2401.08281 (2024)
6. Gilardi, F., Alizadeh, M., Kubli, M.: ChatGPT outperforms crowd workers for text-annotation tasks. Proceedings of the National Academy of Sciences **120**(30), e2305016120 (2023)
7. Grinsztajn, L., Oyallon, E., Varoquaux, G.: Why do tree-based models still outperform deep learning on typical tabular data? Advances in neural information processing systems **35**, 507–520 (2022)
8. OpenAI: Api pricing, https://openai.com/api/pricing/
9. Singh, A., Diao, Y.: Leveraging large language models for optimized item categorization using unspsc taxonomy. arXiv preprint arXiv:2503.04728 (2024)
10. UNDP: United Nations Standard Products and Services Code (UNSPSC), https://www.undp.org/unspsc