Categorization Accuracy Evaluation Methodology for SMB Bookkeeping Models

Prepared by BooksGPT Applied AI Team · Public methodology packet · Version 1.4

This paper describes the evaluation design behind the BooksGPT public benchmark claims for transaction categorization. It specifies the sampling frame, label policy, chart of accounts mapping, baseline comparisons, statistical tests, and release artifacts required to support the figures used in the pricing page and FAQ.

The public page references a 50,000 transaction held-out set across 12 industries, 47 United States banks, and 24 months of anonymized small business activity. This methodology is written so an auditor can reproduce the run from the signed eval export and inspect every accepted label, rejected label, model prediction, confidence score, and review decision.

PUBLIC BENCHMARK CLAIM

91.2%

Categorization accuracy referenced in FAQ and pricing page.

HELD-OUT EVALUATION SET

50K

Anonymized SMB transactions, stratified by industry and bank source.

AUDIT REQUIREMENT

100%

Each public score must map to a signed run manifest and immutable label file.

Abstract

Transaction categorization in small business bookkeeping is a supervised classification task with unusually high stakes for trust. A correct label must reflect the merchant, amount, date, tax treatment, entity history, and the customer chart of accounts. Generic category prediction is not sufficient because two businesses can code the same merchant differently for valid reasons.

This methodology treats the chart of accounts as a first class input. Each model prediction is evaluated against an adjudicated gold label, not against a generic merchant taxonomy. Evaluation records include the raw normalized descriptor, merchant resolution, prior customer history, candidate accounts, selected account, confidence score, verifier decision, fallback route, and reviewer outcome.

> The document is a methodology paper. It does not replace the signed eval export. A public score is considered supportable only when the PDF, manifest, label file, run logs, and missed transaction appendix all reference the same run identifier.

READER GUIDE

WHAT A REAL PAPER SHOULD CONTAIN

A credible methodology paper should look closer to an applied machine learning evaluation report than to a marketing one-pager. It should use tables, label definitions, diagrams, and error analysis. It should not use stock photos. Screenshots are useful only when they show the evaluation harness, the reviewer workflow, or a redacted audit record.

1. Public Claim Registry

The registry reconciles the numbers and phrases used in the pricing page, FAQ, and evaluation link. It is intentionally separate from the raw result table so that each public claim can be traced to the artifact that supports it.

CLAIM SURFACE	CURRENT PUBLIC VALUE	REQUIRED SUPPORTING ARTIFACT	ACCEPTANCE RULE
Pricing benchmark bar	BooksGPT v1.4 categorization accuracy: 91.2%	Signed run manifest, prediction file, adjudicated labels, exact miss file.	Accuracy, macro F1, and Wilson interval calculated from same run ID.
FAQ accuracy statement	91.2% on 50,000 transaction test set.	Held-out set manifest with 50,000 unique row IDs and dedupe hash.	No transaction from training or prompt examples can share a dedupe key.
Benchmark context	12 industries, 24 months, 47 United States banks.	Stratification report with counts per industry, month, and source bank.	Each stratum must meet the minimum count or be disclosed as sparse.
Comparison bar	Senior bookkeeper: 84.1%, GPT-4o zero-shot: 82.3%, QuickBooks auto-rules: 70.4%.	Baseline run notebooks, prompt files, rule definitions, reviewer rubric.	Baselines must run on the same held-out rows and same gold labels.
Audit trail claim	100% of entries have an activity log and reversible posting.	Ledger event schema, posting log sample, reversal test report.	Every generated entry must have actor, timestamp, source, and prior state.

WHY THIS REGISTRY MATTERS

Accuracy claims are easy to misread when the unit of measurement changes. A merchant-level score, transaction-level score, account-level score, top-k score, and reviewer-assisted score are not interchangeable. The public value must name the unit: one final account assignment per transaction, scored against an adjudicated chart of accounts label.

The model score is not accepted from a dashboard alone. It is accepted from an immutable bundle: input rows, predictions, gold labels, config, model version, prompt version, verifier version, and code commit. The bundle should produce the same score when run by a reviewer who has no access to training data.

2. Evaluation Scope and Exclusions

The v1.4 evaluation focuses on one decision: choosing the correct bookkeeping account for an imported bank or card transaction. The transaction can include a merchant descriptor, amount, date, account source, prior categorization history, customer chart of accounts, and industry profile.

INCLUDED DECISIONS

- Expense, income, transfer, liability, asset, equity, and contra account selection.
- Recurring merchant handling when historical customer data is available.
- Cold-start handling when only industry and descriptor context are available.
- Split recommendations when the transaction requires multiple accounts.
- Confidence thresholding, human review routing, and reversible posting.

EXCLUDED DECISIONS

- Tax return preparation and legal interpretation.
- Sales tax nexus decisions, payroll filings, and entity classification.
- Receipt OCR extraction accuracy.
- Fraud detection performance beyond anomaly flag routing.
- User satisfaction, time saved, and support burden.

PRIMARY TASK DEFINITION

For each held-out transaction row **t**, the system receives the available production-time inputs and returns a final account assignment **y_hat**. The gold label **y** is the adjudicated account that an expert reviewer accepts for the specific business chart of accounts. The prediction is correct when **y_hat = y** after canonical account ID mapping.

SECONDARY TASK DEFINITIONS

TASK	OUTCOME	REASON TO MEASURE
Top-3 account recall	Gold account appears in first three candidates.	Measures assistant value when user must pick from options.
Review routing precision	Low-confidence or ambiguous rows go to review.	Prevents confident wrong posts from reaching the ledger.
Split detection	System marks transactions that need multiple accounts.	Captures Stripe, Square, payroll, loan, and owner draw edge cases.
Reversal safety	Every automated post can be reversed with prior state intact.	Connects model quality to bookkeeping control quality.

3. Dataset Construction and Privacy Controls

The held-out set is designed to represent the messy distribution of real SMB bank feeds without exposing raw customer records. The unit of sampling is a transaction after account connection, dedupe, merchant normalization, and personally identifiable information redaction.

SAMPLING FRAME

Transactions imported through bank and card connectors, spanning the period named in the run manifest.

STRATA

Industry, source bank, transaction type, month, amount band, merchant recurrence, and chart complexity.

HOLDOUT

Rows and near-duplicates are excluded from training, prompt examples, few-shot libraries, and retrieval memories.

INCLUSION AND EXCLUSION RULES

RULE	INCLUDED	EXCLUDED
Transaction state	Posted bank and card transactions with stable amount and date.	Pending, reversed before posting, or connector test rows.
Label availability	Rows with final customer accepted category or expert adjudicated label.	Rows still in dispute, rows with missing chart account, rows marked personal only.
Privacy	Redacted descriptors, hashed IDs, generalized notes, amount bands where needed.	Names, addresses, memo fields with private content, tax IDs, account numbers.
Distribution	Common merchants, rare merchants, one-off expenses, transfers, refunds, payroll, card fees.	Data sources not available at production inference time.

DEDUPE AND LEAKAGE CHECKS

Each row receives a dedupe key derived from normalized merchant text, amount bucket, account ID hash, month, and customer hash. A row is excluded if its dedupe key appears in training data, prompt examples, reviewer warmup examples, or any retrieval index exposed to the model during evaluation. The manifest records the matching rule and the exclusion count.

> A high score on leaked recurring merchants is not meaningful. The leakage check is run before every score calculation and is stored with the score output.

4. Gold Labels, Adjudication, and COA Mapping

The gold label is the accepted bookkeeping account in the customer chart of accounts after review. Because charts vary by business, labels are canonicalized twice: first to the customer account ID, then to a cross-company reporting family used for aggregate confusion analysis.

REVIEWER WORKFLOW

1. Masked review
Reviewer sees redacted row, chart options, industry, and allowed context.

2. First label
Reviewer selects account, split flag, and reason code.

3. Conflict check
Disagreement with existing label triggers second reviewer.

4. Adjudication
Senior reviewer resolves conflicts and records final rationale.

CANONICAL COA MAPPING

CUSTOMER ACCOUNT EXAMPLE	CANONICAL FAMILY	TAX POSTURE	AMBIGUITY NOTES
Software, SaaS, subscriptions	Software and online services	Operating expense	May be COGS for software resale business.
Meals, client meals, employee meals	Meals and entertainment	Deduction limited by context	Requires note when event purpose is present.
Owner draw, shareholder distribution	Equity distribution	Not expense	Common confusion with payroll or contractor payment.
Loan payment, principal, interest	Debt service split	Principal and interest split	Requires amortization or lender statement for final split.
Sales tax payable, state tax, marketplace tax	Sales tax liability	Liability	Must not be booked as revenue or expense when collected.

INTER-REVIEWER RELIABILITY

For a public release, at least 10% of the held-out set or 2,000 rows, whichever is larger, should receive dual review. The release packet should report raw agreement, Cohen kappa for account family, and conflict rate by industry. Rows with unresolved disputes are excluded from score calculation and listed in the adjudication appendix.

5. Prompt Graph and Inference Pipeline

The category model is not evaluated as a single isolated prompt. It is evaluated as the same inference graph used in production: normalize the transaction, retrieve relevant context, propose candidate accounts, verify risk, and either post or route for review.

A. Normalize
Clean descriptor, identify merchant, detect transfer markers, map source bank fields.

B. Retrieve
Fetch prior customer handling, industry playbook, account definitions, and merchant memory.

C. Candidate COA
Rank eligible accounts and split candidates before final model call.

D. Categorize
Return account ID, split flag, confidence, evidence, and alternative labels.

E. Verify
Check policy constraints, transfer logic, confidence threshold, and tax sensitive classes.

F. Fallback
Route ambiguous rows to review with reason code and recommended accounts.

G. Post
Write accepted entry to ledger with source fields and rollback pointer.

H. Log
Store prompt version, model version, retrieved IDs, output, verifier decision, and actor.

EVALUATION ISOLATION

The evaluation environment uses the same graph but a frozen model, frozen prompts, frozen retrieval corpus, and read-only ledger target. The model cannot learn from held-out predictions during the run. Human corrections collected after the evaluation date are held for the next training cycle and never backfilled into the score under review.

REQUIRED RUN FIELDS

FIELD	PURPOSE	STORED IN MANIFEST
model_version	Names the exact model or adapter used for categorization.	Yes
prompt_graph_version	Names normalize, retrieve, candidate, model, verifier, and post steps.	Yes
retrieval_snapshot_id	Prevents hidden changes to context after scoring.	Yes
confidence_threshold	Separates auto-post rows from review rows.	Yes
code_commit	Allows rerun from source.	Yes

6. Metrics and Confidence Intervals

The headline score should be simple enough for buyers to understand and strict enough for reviewers to trust. The primary score is transaction-level exact account accuracy. Secondary scores explain where the system is strong, where it is uncertain, and where it fails.

PRIMARY METRIC

Exact account accuracy equals the count of transactions where the final predicted account ID matches the adjudicated gold account ID divided by all scored transactions. Review-routed rows are counted according to the chosen public claim: model-only accuracy counts the pre-review model output, while assisted workflow accuracy counts the accepted final decision after review.

SECONDARY METRICS

- **Macro F1:** balances performance across account families so rare classes matter.
- **Weighted F1:** reflects real transaction volume.
- **Top-3 recall:** measures usefulness when suggestions are shown to a user.
- **False auto-post rate:** wrong predictions that pass the confidence gate.
- **Review burden:** share of rows sent to human review.

PUBLIC SCORE CALCULATION

Accuracy = correct final account assignments / scored transactions.

The public pricing page references 91.2%. The release packet must include the run ID that reproduces the numerator, denominator, and exact miss list.

Macro F1 = unweighted mean of per-family F1 scores.

Wilson CI = binomial confidence interval for exact accuracy.

Bootstrap CI = 10,000 row-resampled intervals for F1 and review burden.

STATISTICAL TESTS

QUESTION	TEST	DECISION RULE
Does v1.4 outperform v1.3 on same rows?	McNemar test on paired correct or incorrect labels.	Report p value and effect size. Do not publish only p value.
Is performance stable across industries?	Bootstrap confidence intervals by industry stratum.	Flag any industry below release floor or with wide interval.
Is review routing safer than auto-posting?	Compare false auto-post rate at threshold candidates.	Select threshold that meets false auto-post ceiling.

7. Baselines and External Context

External studies are useful for calibrating expectations, but they are not substitutes for a BooksGPT held-out run. Published results differ by category granularity, country, bank feed quality, availability of customer history, and whether the model predicts generic classes or customer-specific chart accounts.

SOURCE	TASK AND SETTING	REPORTED FIGURE	HOW TO USE IT HERE
QuickBooks Rel-Cat, 2025	Customer-specific QuickBooks transaction categorization framed as link prediction.	Rel-Cat reported 68.67% Top-1 and 88.04% Top-5 accuracy in few-shot setting.	Shows that flexible customer charts make top-1 account prediction much harder than generic merchant category prediction.
Kreditz collaboration, 2024	Bank transaction categorization using character-level deep learning on selected inflow classes.	Existing engine around 95% accuracy; CNN and LSTM macro F1 reached 0.97.	Useful reference for high precision transaction categorization when class set is constrained.
Bank statement accounting classification, 2025	Matching bank transaction statements to accounting records with FNN and SVM models.	Best encoded-label result: FNN 87.02%, SVM 86.25% on Dataset 1.	Highlights that accounting record matching varies sharply by dataset and label quality.
Gartner finance survey, 2024	Survey of controllership staff about financial errors and technology acceptance.	18% reported daily financial errors; 59% reported several per month; high technology acceptance associated with 75% fewer errors.	Context for workflow control design, not a direct model accuracy baseline.

INTERNAL BASELINES

The pricing page compares BooksGPT v1.4 with a senior bookkeeper workflow, a generic GPT-4o zero-shot prompt without chart of accounts context, and QuickBooks-style auto-rules. For these values to be publishable, every baseline must use the same input rows, the same gold labels, and a frozen protocol. The human baseline should include at least two reviewers and time limits matching the claimed monthly workload.

> Numbers from external studies should never be mixed into the BooksGPT score. They belong in the reference section or a clearly labeled context table.

8. Confusion Matrix and Miss Taxonomy

The confusion matrix is the most important non-headline artifact because it shows whether residual errors are harmless naming differences or material accounting mistakes. The public appendix should include the per-industry matrix, the all-industry aggregate matrix, and the exact missed rows after redaction.

GOLD FAMILY	COGS	SOFTWARE	MEALS	PAYROLL	TRANSFER	TAX LIABILITY	REVIEW
COGS	Run value						
Software		Run value					
Meals			Run value				
Payroll				Run value			
Transfer					Run value		
Tax liability						Run value	

The matrix above shows the published layout. Values are populated from the signed evaluation export so the PDF and scorecard stay tied to the same run.

MISS TAXONOMY

MISS CLASS	DESCRIPTION	ACCOUNTING RISK	EXPECTED HANDLING
Merchant ambiguity	Descriptor does not reveal whether purchase is supplies, meals, software, or inventory.	Medium	Route to review unless prior history is strong.
Account policy mismatch	Two businesses intentionally code the same merchant differently.	Medium	Prefer customer history over generic industry default.
Transfer or payment confusion	Credit card payment, loan payment, owner transfer, or payroll sweep looks like expense.	High	Verifier blocks auto-post and checks counterparty account.
Tax-sensitive split	Sales tax, payroll tax, interest, principal, fees, and tips are mixed in one amount.	High	Require split evidence or reviewer confirmation.
New vendor cold start	Merchant is unseen in customer history and weak in industry memory.	Low to high	Use top candidates and confidence threshold.

9. Release Gate, Audit Trail, and Reproducibility

A model release is not accepted because a single aggregate score improves. It is accepted only if the score improves or holds steady, high-risk misses are below threshold, review routing catches ambiguous rows, and every artifact needed for audit is archived.

RELEASE GATES

1. Exact accuracy meets or exceeds the public threshold for the named release.
2. Macro F1 does not regress by more than the agreed margin.
3. False auto-post rate is below threshold for high-risk account families.
4. No industry stratum falls below the release floor without disclosure.
5. Top miss classes have owner, ticket, and mitigation plan.

AUDIT ARTIFACTS

1. Run manifest and config hash.
2. Held-out row manifest and leakage report.
3. Gold label file and adjudication log.
4. Prediction file, confidence file, and verifier file.
5. Score notebook, PDF, and exact miss appendix.

LEDGER AUDIT TRAIL

Model evaluation is linked to ledger safety through the event log. Each accepted production entry must include source transaction ID, normalized descriptor, model version, prompt graph version, confidence, verifier decision, posting actor, timestamp, and reversal pointer. The evaluation does not claim tax defensibility unless this log exists for every posted entry.

REPRODUCTION COMMAND SHAPE

STEP	INPUT	OUTPUT
validate-holdout	Row manifest, training manifest, retrieval snapshot	Leakage report
run-eval	Model, prompt graph, holdout rows	Prediction file
score-eval	Prediction file, gold labels	Metrics and intervals
export-packet	Metrics, misses, references	Signed release folder

10. Limitations and Monitoring

A held-out benchmark is a release gate, not a permanent guarantee. Bank descriptor formats change, merchants change processors, tax rules change, and a customer's chart of accounts changes as the business grows. The methodology therefore includes online monitoring and periodic refresh.

KNOWN LIMITATIONS

LIMITATION	IMPACT	MITIGATION
Customer-specific bookkeeping policy	Same merchant can be validly coded to different accounts.	Use account definitions, prior history, and reviewer memory.
Rare account families	Aggregate score can hide weak classes.	Report macro F1 and per-family intervals.
Changing bank descriptors	Model can lose merchant context after bank format changes.	Monitor descriptor drift by source bank and month.
Ambiguous tax treatment	Material errors can pass if transaction lacks purpose.	Route tax-sensitive uncertainty to review.
Human label noise	Gold labels can reflect inconsistent bookkeeping decisions.	Dual review, conflict adjudication, and policy notes.

POST-RELEASE MONITORING

- Track customer corrections by account family, bank source, industry, model version, and confidence decile.
- Alert when correction rate rises above release baseline for three consecutive days.
- Sample auto-posted high-confidence rows for human review each week.
- Run a shadow evaluation on new data before changing prompt graph or retrieval policy.
- Publish methodology updates when the task definition, holdout set, baselines, or score calculation changes.

> The score should not be marketed as a guarantee for any particular customer's books. It is a release benchmark over the named held-out distribution and should be read with the limitation table.

11. Publication Packet

The website PDF should be a stable front door into the evidence packet. Buyers can read the plain-language method, while auditors and partners can request the signed artifacts under the appropriate data handling agreement.

FILES IN THE PACKET

FILE	AUDIENCE	CONTENTS	RETENTION
methodology.pdf	Public	Methods, claim registry, metrics, baselines, limitations, references.	Versioned forever.
run-manifest.json	Internal, auditor	Run ID, model version, prompt graph, code commit, dataset hashes.	Versioned forever.
holdout-summary.csv	Internal, auditor	Counts by industry, bank, month, merchant recurrence, account family.	Versioned forever.
misses-redacted.csv	Internal, auditor	Exact missed rows after redaction, with reason code and mitigation owner.	Versioned forever.
score-notebook.html	Internal, auditor	Metric calculations, intervals, tests, and charts.	Versioned forever.

PUBLIC WORDING STANDARD

Public language should name the release, the dataset, and the metric. A strong sentence is: "BooksGPT v1.4 is evaluated on a 50,000 transaction held-out set across 12 SMB industries using exact account accuracy against adjudicated chart of accounts labels." Avoid vague phrases such as "near perfect" or "human-level" unless the comparison protocol is stated.

RECOMMENDED REVIEW BEFORE PUBLISH

ENGINEERING
Confirms run can be reproduced from code and manifest.

ACCOUNTING
Confirms label policy, split handling, and tax-sensitive families.

LEGAL
Confirms public claims match evidence and limitations.

12. References

[1] Kaiwen Dong, Padmaja Jonnalagedda, Xiang Gao, Ayan Acharya, Maria Kissa, Mauricio Flores, Nitesh V. Chawla, and Kamalika Das. "Transaction Categorization with Relational Deep Learning in QuickBooks." arXiv:2506.09234, 2025. https://arxiv.org/abs/2506.09234

[2] Juan Liu, Lei Pei, Ying Sun, Heather Simpson, Jocelyn Lu, and Nhung Ho. "Categorization of Financial Transactions in QuickBooks." Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2021. https://doi.org/10.1145/3447548.3467100

[3] Filip Orestav and Kolumbus Lindh. "Deep Learning Techniques for Bank Transaction Categorization: A Collaborative Study with Kreditz." KTH Royal Institute of Technology, 2024. https://www.diva-portal.org/smash/get/diva2:1884903/FULLTEXT01.pdf

[4] "Accounting Support Using Artificial Intelligence for Bank Statement Classification." Computers, 14(5), 193, 2025. https://www.mdpi.com/2073-431X/14/5/193

[5] Duc Tuyen Ta, Wajdi Ben Saad, and Ji Young Oh. "Specialized Text Classification: An Approach to Classifying Open Banking Transactions." arXiv:2504.12319, 2025. https://arxiv.org/abs/2504.12319

[6] Gartner. "Gartner Survey Shows That a Third of Accountants Make Several Financial Errors Per Week due to Capacity Constraints." Press release, February 21, 2024. https://www.gartner.com/en/newsroom/press-releases/2024-02-21-gartner-survey-shows-that-a-third-of-accountants-make-several-error-per-weeo-due-to-capacity-constraints

[7] QuickBooks Support. "Categorize Online Bank Transactions in QuickBooks Online." Intuit, current support documentation. https://quickbooks.intuit.com/learn-support/en-us/help-article/banking/categorize-match-online-bank-transactions-online/L1bTafTz3_US_en_US

[8] QuickBooks Support. "How AI Suggestions Help Match and Categorize Bank Transactions." Intuit, current support documentation. https://quickbooks.intuit.com/learn-support/en-us/help-article/bank-transactions/ai-suggestions-help-match-categorize-bank/L8FHOh4AD_US_en_US

SOURCE NOTE

External numbers are included as context only. The BooksGPT public values must be supported by the internal run artifacts listed in the publication packet. The references are used to explain why transaction categorization should be evaluated against customer-specific charts, held-out rows, leakage controls, and per-class error analysis.