Neurosymbolic Artificial Intelligence
Revised manuscript
Resubmission
Tracking number 1003-2039
11 September 2026

TaxBridge-Bench: A Neurosymbolic Benchmark for Spreadsheet-to-Ledger Reasoning in Tax Compliance

David Elliman

Abstract

We introduce TaxBridge-Bench, a benchmark for neurosymbolic spreadsheet-to-ledger reasoning in UK Making Tax Digital tax compliance. It reconstructs income and expense totals from workbooks containing values, formulas, formatting, dates, VAT treatment, and trade context. It comprises a 30-case evaluation split with 163 SA103F box values and a 40-case deployment corpus generated from executable specifications. It tests preservation of workbook evidence, monetary aggregation, defeasible correction, and bounded use of large language models (LLMs). An LLM debate panel obtains 107/163 boxes but no case pass. A serialised-workbook Sonnet 5 baseline returns valid JSON for 39/40 deployment cases but achieves 56/259 boxes and 0/40 case passes. A symbolic-first blackboard system passes the aggregate-box scorer on all 30 evaluation cases and the money, period and VAT release checks on all 40 deployment cases. These criteria have different scopes and do not establish full-ledger equivalence. The synthetic results motivate bounded neural reasoning over preserved workbook evidence; generalisation requires independent evaluation and stronger baselines.

Keywords

neurosymbolic AI; benchmark; spreadsheet understanding; document reasoning; tax compliance; blackboard systems

Neuro-Symbolic Ltd, Gloucestershire, United Kingdom

Corresponding author:
David Elliman, Neuro-Symbolic Ltd, Gloucestershire, United Kingdom.
Email: dave@neusym.ai

Prepared using sagej.cls *[Version: 2016/06/24 v1.10]*

Introduction

Benchmarks for spreadsheet reasoning usually ask whether a system can answer questions about, manipulate, or summarise workbooks. Regulated document processing poses a different problem: a system must preserve the evidence needed to compute legally meaningful totals, reject unsupported inferences, and leave an audit trail explaining how each output follows from the source document. This paper presents a benchmark for that setting.

TaxBridge-Bench evaluates quarterly sole-trader tax spreadsheets under the UK Making Tax Digital for Income Tax regime. A system receives an Excel workbook and business context, then must produce HMRC-compatible income and expense totals. Workbooks may contain offset tables, formula totals, split gross/net VAT treatment, colour-coded personal expenses, ambiguous descriptions, and fiscal-boundary rows. A visible-cell serialisation omits some of this evidence.

The benchmark is intended for neurosymbolic systems because successful solutions require both symbolic and neural capabilities. Symbolic components are well suited to dates, formulas, amounts, row inclusion, arithmetic constraints, and audit trails. Neural components are useful for ambiguous headers, supplier normalisation, and trade-specific semantic judgement. The research question is where neural reasoning should sit within an exact and inspectable workflow.

The reference architecture and the blind-split and diagnostic-corpus results were previously described in an architecture-focused preprint (Elliman, 2026a). The present manuscript develops the benchmark task, metrics, illustrative cases, and reproducibility requirements, and adds the 40-case deployment corpus and serialised-workbook baseline. It contributes:

1. A document-to-ledger task that exposes structural spreadsheet evidence usually lost by text-only LLM baselines.
2. Two reusable synthetic corpora: a 30-case blind adversarial split and a 40-case executable go/no-go corpus generated from `.spec.json` fixtures.
3. Strict scoring metrics for exact ledger reconstruction, SA103F box accuracy, row inclusion, VAT treatment, period handling, and contextual flags.
4. Baseline results for an LLM-first debate panel, a simple serialised-workbook LLM, an author-confirmed AutoGen-style pilot, and a symbolic-first blackboard reference system.
5. A reproducibility specification covering workbooks, ground truth, prompts, outputs, scoring reports, and harnesses.

Benchmark Task

Input and Output

Each case contains an Excel workbook and a small context record specifying the tax year, update period, VAT status where known, and accounting basis. The target output is a structured ledger suitable for HMRC quarterly update computation. At minimum,

a system must return the official income and expense box totals, identify excluded or personal rows, respect the fiscal period, and preserve VAT net/gross treatment.

Workbooks represent ten trades, including electricians, builders, hairdressers, couriers, market traders, consultants, photographers, cleaners, plumbers, and driving instructors. Layouts range from cashbooks to bank exports with noisy supplier descriptions.

Illustrative Case

Table 1 shows a representative non-VAT plumber case from the deployment corpus. It is simple enough for an accountant to read, but already contains the three properties that make the benchmark non-trivial: a fiscal-boundary row visible in the workbook but excluded from Q1 totals, ordinary trade expenses that require category mapping, and a personal drawing that must remain visible while contributing zero to claimable expenses. The reference system retains row position, sheet membership, period context, status annotations, and arithmetic constraints as typed evidence.

Table 1. Representative rows from a Q1 non-VAT plumber fixture.

Workbook row	Relevant evidence	Ground-truth treatment
3 Apr customer payment	Standard-period Q1 starts on 6 Apr	Turnover category, but excluded from Q1 totals
Screwfix copper fittings	Supplier, trade, and raw "Materials" label	Included as cost of goods
Mobile phone business share	Raw "Admin" label conflicts with description	Included as phone and internet
Owner drawings to personal account	Expense-looking row with personal semantics	Visible personal row, claimable amount zero

Neurosymbolic Properties Tested

The benchmark targets four properties of neurosymbolic systems.

Multi-modal workbook evidence. Correct results often depend on evidence outside the visible text stream: cell location, table boundaries, formulas, fills, and implicit section structure.

Bounded neural reasoning. LLM calls should resolve ambiguity without being allowed to invent amounts, silently change row counts, or override deterministic accounting constraints.

Defeasible correction. A later validator may contradict or supersede an earlier classification. The scorer rewards final correctness, while the artefacts preserve enough provenance to inspect how disagreements were resolved.

Auditable aggregation. Box totals must trace back to source rows and pass pence-level arithmetic checks. Near-correct summaries are not enough for regulatory submission.

Dataset and Artefacts

TaxBridge-Bench Blind Split

The evaluation split, called the blind split in the experiment records, contains 30 synthetic workbooks, cases 031–060, 163 SA103F box values, and 9 VAT-registered cases. The reference system emits 489 transactions. Recovered records contain successive runs on this split and do not establish that it remained untouched during development. Adversarial features include:

- random table origin offsets rather than standardised cell positions;
- native `SUM` formulas rather than hardcoded aggregate values;
- background fill colours marking personal or suspicious entries;
- out-of-period rows around quarter boundaries;
- VAT standard-scheme and flat-rate cases requiring net/gross preservation;
- mixed-use expenses and drawings/wages ambiguity.

Deployment-Readiness Go/No-Go Corpus

The second corpus contains 40 generated workbooks and 1,069 rows. Each workbook is generated from a runnable `.spec.json` fixture containing workbook layout, business context, row-level `hmrc_truth`, and expected current-engine interpretation. The corpus provides broader synthetic coverage than the blind split: 10 cases per quarter, 16 non-VAT cases, 16 standard-VAT cases, 8 flat-rate VAT cases, 20 cash-basis cases, 20 traditional-basis cases, and 10 trades across all four quarters.

The go/no-go corpus is useful for repeatable regression testing because the source specifications can regenerate the spreadsheets and ground truth. It also supports a simple LLM baseline: each workbook can be serialised to visible non-empty cells and submitted without access to the specification or scorer.

Case Construction and Controls

Each executable specification records the tax year, quarter, accounting basis, VAT status and scheme, period convention, business name, and whether displayed amounts include VAT. It specifies workbook layout and rows, with separate `hmrc_truth` and `expected_current_engine` fields. Target JSON records context, workbook shape, box totals, row truth, and flags.

Regeneration permits inspection of each workbook against its specification. Row-level truth exposes errors that cancel at aggregate level. Separating expected engine output from the intended HMRC treatment makes implementation limitations visible.

The generator varies five axes deliberately: trade profile, VAT scheme, accounting basis, update-period convention, and layout style. Layouts include combined cashbooks, split income/expense sheets, quarterly tabs, and bank-export-style sheets. Edge rows are injected around period boundaries: 1–5 April rows distinguish standard from calendar periods, immediately-after-quarter rows test cumulative update semantics, and after-tax-year rows test exclusion at the annual boundary. Other rows test drawings, vehicle

purchases, capital treatment under cash versus traditional basis, exempt or zero-rated VAT treatment, flat-rate VAT capital goods, and trade-specific supplier descriptions.

The two corpora have different evidential roles. The blind split reports retrospective evaluation outcomes; the deployment corpus tests the production path across trades and quarters. The latter is a regression suite, not an independent held-out evaluation.

Panel Corpus

A third 45-case synthetic panel corpus captures further layout variation, including Monzo- and Starling-style bank exports. It supports diagnostic analysis of supplier normalisation and human-review edge cases. Its historical comparison records are separate from the two main evaluation corpora.

Metrics

The intended strict metric is **exact ledger reconstruction**: every required box total must match within £0.01 and all row checks must pass. The recovered experiments use narrower, distinct scorers. The blind scorer checks expected aggregate boxes, grouping office costs with other expenses; it does not independently verify all rows or reject every extra box. Its case-pass rate therefore measures agreement under that scorer, rather than complete ledger correctness.

We also report partial metrics:

- **Box accuracy**: number of correct SA103F or MTD box totals divided by the number of ground-truth totals.
- **Output transaction count**: rows emitted by a system; counts alone do not establish row-by-row recovery.
- **VAT-case pass rate**: aggregate-box scorer passes on VAT-registered cases.
- **Flag accuracy**: contextual flags such as accounting basis, VAT scheme, period type, and quality conditions.
- **Cost and latency**: LLM tokens and wall-clock time where available.

Missing responses remain in the baseline denominators. The simple baseline scores named boxes separately. Its flag scorer uses literal string matching, so alternative labels such as “standard and “standard_vat can count as errors despite similar meaning.

The deployment specification runner checks row inclusion, claimable amounts, period membership, VAT treatment and summary totals. Category-box and colour differences are advisory and do not prevent a release pass. The separate app-path gate checks income, expense and net totals through the upload-to-draft workflow. Neither gate establishes correctness of every category box; their pass rates are not directly comparable with the simple baseline’s box accuracy.

Unresolved rows may be represented as discrepancies, but they do not count as correct unless their monetary treatment is resolved. Omitted boxes remain in the denominator: referral for human review is a safe failure rather than a completed update.

Token use and latency are secondary metrics. Reported plus/minus values are sample standard deviations across cases. Measurements from separate runs are not a controlled cost comparison.

Systems Evaluated

LLM Debate Panel

The first baseline is an LLM-first debate panel: four coordinated roles (semantic proposer, quantitative proposer, critic, and chair) receive a serialised representation of the workbook and debate toward a consensus ledger. This was a serious domain-tuned baseline rather than a deliberately weak prompt. Its observed errors are consistent with loss of workbook evidence during serialisation. The comparison also changes architecture and tool access, so it does not isolate serialisation as the sole cause.

Simple Serialised-Workbook LLM

The second baseline serialises each workbook to visible non-empty cells and asks `claude-sonnet-5` to emit JSON totals and flags. It has no workbook parser, no tools, and no access to the specifications or ground truth. This baseline is intentionally simple because it measures how much signal survives ordinary text serialisation.

Earlier AutoGen-Style Pilot

An earlier AutoGen-style orchestration pilot used the same broad LLM-first strategy: a serialised workbook representation was passed through conversational agents rather than preserved as typed workbook evidence. The author has confirmed that this pilot obtained approximately 17–18% box/category accuracy under the earlier scoring setup. Its raw prompts, configuration, and outputs are still being recovered, so Table 3 reports it as an artefacts-pending pilot rather than as a headline reproducible result.

Symbolic-First Blackboard Reference System

The reference system is a neurosymbolic blackboard. Deterministic Knowledge Sources ingest cells, detect dates and amounts, assemble tables, evaluate fiscal-period membership, preserve formulas and fill colours, classify obvious suppliers, handle VAT, and validate totals. LLM-backed Knowledge Sources are gated behind symbolic uncertainty: they interpret ambiguous headers, normalise suppliers, and arbitrate a small number of flagged boundary cases. Persistent hypotheses are immutable and can be superseded only by later hypotheses that preserve provenance.

This reference system is included to demonstrate that the benchmark is solvable without sending the entire problem to an LLM. It is not presented as the only possible architecture.

Results

Blind Split

Table 2 reports the saved comparison of 27 March 2026 at 19:24:34. The blackboard passes the aggregate-box scorer on every case; the debate panel obtains many individual boxes but no case pass.

Table 2. TaxBridge-Bench blind split, cases 031–060.

Metric	LLM debate panel	Blackboard reference
Cases passing box scorer	0/30 (0%)	**30/30 (100%)**
SA103F box accuracy	107/163 (66%)	**163/163 (100%)**
Output transactions	399/489 (82%)	**489/489 (100%)**
VAT cases passing box scorer	0/9	**9/9**
Mean latency per case	158.8 s ± 29.4	**6.2 s ± 1.1**
LLM tokens per case	37,788 ± 3,391	Not recorded for full run
Total LLM tokens	≈1,134k	Not recorded for full run

The debate panel achieves 66% box accuracy under the historical grouping rule, but no case passes all expected-box checks. Observed errors are consistent with omitted formula structure, fill colours, and row location; the experiments do not isolate the contribution of each feature.

Deployment-Readiness Corpus

Table 3 combines the 18 August 2026 app-path and specification-runner records with a saved simple-LLM baseline. The baseline receives visible-cell serialisations of the same workbooks. The 1,069 rows are cumulative quarterly row occurrences, not 1,069 distinct transactions.

Table 3. Deployment-readiness corpus, simple LLM baseline, and earlier AutoGen-style pilot.

Metric	Value
Cases / rows	40 / 1,069
MTD box totals in ground truth	259
Context and quality flags	280
VAT coverage	16 non-VAT, 16 standard VAT, 8 flat-rate VAT
Accounting basis coverage	20 cash, 20 traditional
App-path cases passed	40/40
Release-clean rows	1,069/1,069
Mean app-path latency	11.9 s ± 7.1
Simple LLM valid JSON responses	39/40
Simple LLM exact cases	0/40
Simple LLM box accuracy	56/259 (21.6%)
Earlier AutoGen-style pilot[a]	approx. 17–18% box/category accuracy; raw artefacts pending
Simple LLM flag accuracy	178/280 (63.6%)
Simple LLM VAT-registered flag accuracy	39/40 (97.5%)
Simple LLM tokens	192,672 total (≈4,817/case)
Simple LLM latency	24.5 s ± 16.8

[a] Author-reported pilot under different scoring. Prompts, configuration, outputs and scorer report are unavailable; this approximate figure is not directly comparable with deployment-corpus scores.

The simple LLM often recognises VAT registration, while period, drawings, VAT and aggregation errors prevent a case pass. The 40/40 app-path and row-release passes are narrower money/period/VAT results; category differences remained advisory.

Diagnostic Corpus

A separate 45-workbook corpus provides diagnostic coverage of spreadsheet layouts. Its manifest and targets have been recovered. The available 26 March comparison record differs from the diagnostic run described in the earlier preprint, so those earlier quantitative diagnostic claims are not used as verified results here.

Observed Error Modes

Table 4 summarises the recurring failure modes observed across the LLM-first baselines. These errors concern structural evidence and accounting treatment as well as recognition of individual categories.

Table 4. Challenge types targeted by the benchmark and observed in baseline errors.

Challenge	Typical evidence	Why serialised LLMs fail
Fiscal boundary	Date plus update-period convention	The string date is visible, but the cumulative MTD period rule is not reliably applied
VAT net/gross	VAT status, scheme, and row treatment	Gross amounts look arithmetically valid even when the ledger requires net values
Visual status	Fill colour or annotation marks personal, warning, or excluded rows	Formatting is lost or demoted to prose during serialisation
Layout semantics	Split sheets, quarterly tabs, offset tables, formula totals	Table membership and formulas are flattened into a single text stream
Trade-specific suppliers	Bank-mangled descriptions such as supplier references and card strings	Semantic normalisation helps, but only if linked back to exact source rows

Diagnostic Ablation

A proposed ablation disables discrepancy-posting validators to test a single forward pass over the same Knowledge Sources. The previously described 100%-to-82% personal-expense result lacks a recovered case denominator and matching outputs, so it is not treated as verified ablation evidence.

Discussion

What the Benchmark Shows

The benchmark isolates a failure mode that is easy to miss in ordinary spreadsheet QA: text serialisation can preserve enough information for plausible partial answers while destroying the evidence needed for exact financial reconstruction. A system can obtain many correct boxes while failing every case under the aggregate scorer. Full ledger correctness needs stronger row-level validation.

The results motivate preserving workbook evidence, bounding neural semantic decisions, and reporting exact-case accuracy alongside partial scores. Because the compared systems differ in representation, tools, and architecture, the contribution of each design choice remains to be established by controlled experiments.

Security and Prompt Injection

The benchmark does not establish resistance to prompt injection. The reference system limits LLM exposure to ambiguous rows, validates responses against category enumerations, rejects invented amounts, and checks totals. These controls require separate adversarial evaluation against spreadsheet-carried instructions.

Why Priority Scheduling Is Used

The reference system uses deterministic priorities to support reproducibility and regression testing. The benchmark also permits volunteer routing, where agents decide whether to act on a request; scoring depends on final correctness rather than scheduling policy.

What Counts as a Stronger Future Baseline

The intended comparison class is not limited to blackboard systems. A stronger future baseline could use a vision-language model over rendered sheets, an agent framework with spreadsheet tools, a program synthesis step that reconstructs tables before prompting, or a hybrid parser with LLM category judgement. Such systems should be evaluated under the same constraints: no access to `.spec.json` truth, pence-exact totals, row-level provenance, and explicit handling of unresolved rows. Independent comparison requires public access to the scorer and fixture generator.

Limitations

Synthetic data. The corpora are synthetic, although designed from observed spreadsheet patterns. No raw customer data is included. Results on production traffic may differ.

Single regulatory domain. The benchmark tests UK tax-compliance spreadsheets. It should not be treated as evidence that the reference architecture will transfer unchanged to medicine, fault diagnosis, or other regulated domains.

Baseline coverage. The evaluated LLM-first systems are a production debate panel and a simple serialised-workbook Sonnet 5 baseline. An earlier AutoGen-style

pilot is reported as an author-confirmed, artefacts-pending result because the prompt, configuration, and raw outputs have not yet been located in a form suitable for archival citation. We have not yet implemented traceable LangGraph, SheetAgent-style, or vision-language baselines with equivalent tool access.

Token accounting. Debate-panel tokens were recorded for the complete blind run. The recovered blackboard token probes contain only case 031 (1,291 and 1,321 tokens in separate runs); zero token fields in the full-run record indicate missing instrumentation. They do not support a corpus-wide token mean or total, which are therefore omitted.

Artefact coverage. The versioned archive provides the datasets, scorer snapshots and recovered experimental records. It does not include the complete production runtime or all historical model outputs. The earlier AutoGen-style pilot remains outside the reproducible comparison because its raw artefacts are unavailable.

Reproducibility Requirements

Reproducing the benchmark requires generated workbooks, source `.spec.json` fixtures where applicable, row-level target JSON, box totals, contextual flags, and workbook-shape metadata. The scoring implementation must document its monetary tolerance, dynamic denominators, equivalent-box rules, and release checks. Baseline records must identify prompts, input serialisations, model versions, inference settings, raw outputs, token measurements, and per-case scores.

A release manifest must identify the benchmark version and date, supported tax year, licences, and split usage. Workbooks must remain inspectable as Excel files so that a reported total can be traced to source rows. The specifications and ground truth are evaluation artefacts and must not be exposed to systems during inference. Failed cases and known implementation limitations are necessary for interpreting the reported scores.

Related Work

Classical blackboard systems, beginning with Hearsay-II and Nii's surveys, showed how heterogeneous Knowledge Sources can cooperate over a shared evidence store (Erman et al., 1980; Nii, 1986a,b). Truth-maintenance systems provide the background for defeasible revision and dependency tracking (de Kleer, 1986; Doyle, 1979). Recent LLM blackboard work revives the pattern for agent coordination, often keeping LLMs as the primary reasoners (Han & Zhang, 2025; Salemi et al., 2026). TaxBridge-Bench instead asks when symbolic structure should dominate and where neural methods should be allowed to intervene.

SpreadsheetLLM, SheetAgent, and vision-language spreadsheet work show that spreadsheet reasoning is an active LLM research area (Chen et al., 2025; Dong et al., 2025; Xia et al., 2024). Their emphasis is representation and general spreadsheet reasoning; SpreadsheetLLM explicitly encodes structural and formatting information, so text serialisation need not discard all workbook structure. The present benchmark differs by focusing on regulated document-to-ledger reconstruction, where exact arithmetic,

source-row provenance, and legally meaningful category boundaries are the scoring target.

Conclusion

TaxBridge-Bench evaluates systems combining symbolic spreadsheet parsing, bounded neural semantic judgement, defeasible correction and auditable aggregation. The reference system passes the recorded blind box scorer and deployment release gates; the LLM-first baselines obtain partial scores without a case pass. Different scoring scopes limit direct comparison. The results motivate preserving structural evidence in financial document reconstruction. Future work should add stronger tool-using baselines, prompt-injection tests and independently contributed systems.

Acknowledgements

OpenAI Codex assisted with manuscript revision, reference checking, the audit of saved experimental records, and preparation of the resubmission materials. The author is responsible for the manuscript's content.

Statements and Declarations

Ethical considerations

Ethical approval was not required. The study uses synthetic spreadsheet data and involves no human participants, human tissue, or raw customer records.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Declaration of conflicting interest

David Elliman is the founder of Neuro-Symbolic Ltd and the developer of the TaxBridge reference implementation evaluated in this study.

Funding

The research was funded from Neuro-Symbolic Ltd's own resources. No external funding was received.

Data availability

TaxBridge-Bench version 1.0.0 (11 September 2026) is openly available on Zenodo at https://doi.org/10.5281/zenodo.22707695 (Elliman, 2026b). The archive contains 115 synthetic Excel workbooks (30 evaluation, 40 deployment and 45 diagnostic), ground truth, deployment specifications and row targets, scorer/harness source snapshots, saved simple-LLM prompts and responses, recovered comparison records, and a checksum manifest. Data and documentation are licensed under CC BY 4.0; the eight included source-code files are licensed under MIT, with file-level scope documented in the archive. The complete production reference runtime and original blind-run prompts/raw responses are not included. Matching original diagnostic, ablation and AutoGen-pilot records remain incomplete.

References

Chen, Y., Yuan, Y., Zhang, Z., Zheng, Y., Liu, J., Ni, F., Hao, J., Mao, H., & Zhang, F. (2025). SheetAgent: Towards a generalist agent for spreadsheet reasoning and manipulation via large language models. In *Proceedings of the ACM on Web Conference 2025* (pp. 158–177). Association for Computing Machinery. https://doi.org/10.1145/3696410.3714962

de Kleer, J. (1986). An assumption-based TMS. *Artificial Intelligence, 28*(2), 127–162. https://doi.org/10.1016/0004-3702(86)90080-9

Dong, H., Zhao, J., Tian, Y., Xiong, J., Xia, S., Zhou, M., Lin, Y., Cambronero, J., He, Y., Han, S., & Zhang, D. (2025). *SpreadsheetLLM: Encoding spreadsheets for large language models* (Version 2) [Preprint]. arXiv. https://arxiv.org/abs/2407.09025v2

Doyle, J. (1979). A truth maintenance system. *Artificial Intelligence, 12*(3), 231–272. https://doi.org/10.1016/0004-3702(79)90008-0

Elliman, D. (2026a). *Bounded LLM reasoning: A neuro-symbolic blackboard architecture for production document analysis* (Version 1.1) [Preprint]. Zenodo. https://doi.org/10.5281/zenodo.21426202

Elliman, D. (2026b). *TaxBridge-Bench: Synthetic workbooks, ground truth, scorers and evaluation records* (Version 1.0.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.22707695

Erman, L. D., Hayes-Roth, F., Lesser, V. R., & Reddy, D. R. (1980). The Hearsay-II speech-understanding system: Integrating knowledge to resolve uncertainty. *ACM Computing Surveys, 12*(2), 213–253. https://doi.org/10.1145/356810.356816

Han, B., & Zhang, S. (2025). *Exploring advanced LLM multi-agent systems based on blackboard architecture* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2507.01701

Nii, H. P. (1986a). Part one: The blackboard model of problem solving and the evolution of blackboard architectures. *AI Magazine, 7*(2), 38–53. https://doi.org/10.1609/aimag.v7i2.537

Nii, H. P. (1986b). Part two: Blackboard application systems and a knowledge engineering perspective. *AI Magazine, 7*(3), 82–107. https://doi.org/10.1609/aimag.v7i3.550

Salemi, A., Parmar, M., Goyal, P., Song, Y., Yoon, J., Zamani, H., Pfister, T., & Palangi, H. (2026). *LLM-based multi-agent blackboard system for information discovery in data science* (Version 2) [Preprint]. arXiv. https://arxiv.org/abs/2510.01285v2

Xia, S., Xiong, J., Dong, H., Zhao, J., Tian, Y., Zhou, M., He, Y., Han, S., & Zhang, D. (2024). Vision language models for spreadsheet understanding: Challenges and opportunities. In J. Gu, T.-J. Fu, D. Hudson, A. Celikyilmaz, & W. Wang (Eds.), *Proceedings of the 3rd Workshop on Advances in Language and Vision Research* (pp. 116–128). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.alvr-1.10