Neurosymbolic Artificial Intelligence
XX(X):1–12

DOI: 10.1177/ToBeAssigned
www.sagepub.com/

SAGE

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

Anonymous author(s)

Abstract

We introduce TaxBridge-Bench, a benchmark for neurosymbolic document reasoning in UK Making Tax Digital tax compliance. The task is to reconstruct HMRC ledger totals from messy sole-trader spreadsheets whose relevant evidence is distributed across cell values, spatial layout, formulas, formatting, dates, VAT treatment, and trade context. The benchmark contains a 30-case blind adversarial split with 489 ground-truth transactions and 163 SA103F box values, plus a 40-case deployment-readiness corpus generated from executable spreadsheet specifications. It evaluates properties under-served by existing spreadsheet benchmarks: preservation of non-textual workbook evidence, strict monetary aggregation, defeasible correction of earlier hypotheses, and bounded use of LLMs inside auditable symbolic workflows. We report LLM-first baselines, an author-confirmed AutoGen-style pilot, and a symbolic-first reference result. A four-role LLM debate panel achieves 107/163 box accuracy but 0/30 exact ledger cases. A fresh serialised-workbook Sonnet 5 baseline on the deployment corpus returns valid JSON for 39/40 cases but obtains 56/259 box accuracy and 0/40 exact cases. A symbolic-first blackboard reference system achieves 30/30 exact cases on the blind split and passes 40/40 deployment cases.

Keywords

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

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. The task is difficult because the relevant evidence is not only textual. Some workbooks use offset tables, formula totals, split gross/net VAT treatment, colour-coded personal expenses, ambiguous supplier descriptions, and rows near fiscal-period boundaries. Serialising the sheet to text destroys much 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 central research question is not whether an LLM can read a spreadsheet, but where neural reasoning should sit in a system that must be exact, bounded, and inspectable.

This manuscript is framed as a benchmark paper. 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 package design covering workbooks, specifications, ground truth, prompts, raw outputs, scorer 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.

The benchmark is deliberately not a template-filling exercise. Workbooks are plausible sole-trader cashbooks rather than clean accounting exports. They include electricians, builders, hairdressers, couriers, market traders, consultants, photographers, cleaners, plumbers, and driving instructors. Some are cashbook-style sheets; others resemble bank

exports where supplier names are noisy and descriptions must be normalised before category matching.

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. A text-only model usually sees the same strings as the reference system. The difference is that the reference system also keeps row position, sheet membership, period context, colour/status annotations, and arithmetic constraints as typed evidence rather than as optional prompt text.

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 primary benchmark split consists of 30 held-out synthetic workbooks, cases 031–060, generated after the reference architecture was frozen. The split contains 489 ground-truth transactions, 163 SA103F box values, and 9 VAT-registered cases. 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 external validity 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

The executable specifications are part of the benchmark rather than merely a convenient test generator. Each specification contains a `config` block for tax-year, quarter, accounting basis, VAT registration and scheme, update-period convention, business name, and whether displayed amounts include VAT. It also contains the workbook header, layout description, row list, and for every source row both `hmrc_truth` and `expected_current_engine`. The generated target JSON separates discovery context, obligation context, workbook shape, MTD box totals, row-level transaction truth, and expected contextual flags.

This structure gives the benchmark three useful properties. First, the workbook can be regenerated, so reviewers can inspect whether the spreadsheet really encodes the claimed task. Second, the ground truth is row-level rather than only aggregate-level, so a system cannot obtain a correct total by cancelling two wrong decisions. Third, the expected engine output is stored separately from the normative HMRC truth, making known implementation limitations visible rather than silently redefining the task.

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 blind split follows a stricter split discipline. Cases 031–060 were generated after the reference architecture was frozen and were scored without tuning against their outcomes. Their purpose is adversarial structural recovery: offset table origins, native formulas, visual metadata, and VAT rows that require preserving gross/net relationships. The deployment corpus has a different role: it is a release-readiness suite that exercises the production path and documents wider coverage across trades and quarters. Treating the two corpora separately prevents the paper from presenting all tests as if they had the same evidential status.

Panel Corpus

A third 45-case synthetic panel corpus captures additional real-world spreadsheet diversity, including Monzo- and Starling-style bank exports. It was used for diagnostic analysis and ablation, not as a comparative benchmark split. It is included because it illustrates the kinds of supplier normalisation and human-review edge cases that motivated the benchmark.

Metrics

The strictest metric is **exact ledger reconstruction**: a case is correct only if every required box total matches ground truth within £0.01 and all release-blocking row checks pass. This all-or-nothing metric is intentionally severe. In tax filing, a ledger with six correct boxes and one wrong box is still not submission-ready.

We also report partial metrics:

- **Box accuracy**: number of correct SA103F or MTD box totals divided by the number of ground-truth totals.
- **Transaction recovery**: source rows represented in the reconstructed ledger.
- **VAT-case correctness**: exact correctness 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.

The scorer uses dynamic denominators so that a system cannot improve its score by omitting difficult boxes. A small set of professionally interchangeable box pairs can be treated as equivalent when they do not change claimable totals; release-blocking checks remain pence-exact.

The deployment scorer adds row-level release checks. For each transaction it compares inclusion/exclusion, income versus expense polarity, HMRC category, claimable amount, VAT treatment where applicable, and fiscal-period status. Summary totals are then

recomputed from the accepted rows rather than trusted from model output. This is important because spreadsheet-to-ledger systems can fail in two different ways: they can misread a source row, or they can read rows correctly but aggregate them into the wrong box. The benchmark records both.

The scorer is intentionally asymmetric with respect to uncertainty. A system is not rewarded for refusing to classify ordinary rows, because omitted boxes remain part of the denominator. Conversely, rows that genuinely require human review can be represented as discrepancies, but they do not become correct totals unless their monetary treatment is resolved. This reflects the compliance setting: safe failure is preferable to silent misfiling, but a safe failure is still not a completed quarterly update.

For cost metrics, token and latency figures are secondary evidence. They are reported because LLM-first systems can be economically impractical even when partially accurate, but they are not used to decide correctness. Where token runs were not collected under identical instrumentation, the paper marks the comparison as approximate rather than treating it as a controlled systems benchmark.

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 failure mode is structural: the serialised input loses workbook evidence needed for several tax boundary conditions.

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 blind-split comparison. The blackboard reconstructs every held-out ledger exactly, while the LLM debate panel obtains many individual boxes but no fully correct case.

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

Metric	LLM debate panel	Blackboard reference
Exact ledger cases	0/30 (0%)	**30/30 (100%)**
SA103F box accuracy	107/163 (66%)	**163/163 (100%)**
Transactions recovered	399/489 (82%)	**489/489 (100%)**
VAT-registered cases exact	0/9	**9/9**
Mean latency per case	158.8 s $\pm$ 29.4	**6.2 s $\pm$ 1.1**
LLM tokens per case	37,788 $\pm$ 3,391	$\approx$**1,300**
Total LLM tokens	$\approx$1,134k	$\approx$**39k**

The 0/30 result for the debate panel should not be read as 30 empty failures. It achieved 66% box accuracy, but strict ledger scoring exposes the practical problem: a single wrong or missing box makes a case not submission-ready. The strongest error pattern is serialisation loss. Once the input representation omits formula structure, fill colours, or row location, the debate has no reliable path back to those facts.

Deployment-Readiness Corpus

Table 3 reports the go/no-go corpus. The app-path run exercises workbook upload, server-side parsing, blackboard analysis, and draft-total generation. The simple LLM baseline uses the same workbooks but receives only a visible-cell serialisation.

These results are useful because they prevent an overly simple interpretation of the 0% panel exact-case result. The LLM-first systems are not hopeless: they recognise some categories and usually recover contextual facts such as VAT registration. They fail where serialisation is weakest: period boundaries, drawings and personal treatment, cumulative quarter semantics, VAT/accounting-basis interactions, and category totals that require structural row evidence.

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 $\pm$ 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 ($\approx$4,817/case)
Simple LLM latency	24.5 s $\pm$ 16.8

[a] Author-confirmed prior pilot. It is included to show that an AutoGen-style LLM-first orchestration produced the same low-accuracy regime as the fresh serialised-workbook baseline, but it should be promoted to a primary baseline only after the prompts, configuration, raw outputs, and scorer report are archived.

Diagnostic Corpus

On the 45-case panel corpus, the blackboard classified 269/270 transactions (99.6%), used 91,933 LLM tokens in total ($\approx$2,043 per case), and averaged 7.7 s per case. The supplier-normalisation loop fired on 17/45 cases (38%), concentrated in messy bank-export-style workbooks. The single unclassified transaction was a multi-platform delivery-driver edge case requiring human review. This corpus is diagnostic, not comparative; no LLM-first baseline was run on it.

Observed Error Modes

Table 4 summarises the recurring failure modes observed across the LLM-first baselines. The point is not that current models cannot identify a plumber's materials purchase or a hairdresser's salon supplies. They often can. The failure comes from treating the workbook as a bag of strings rather than as a structured source document whose non-textual evidence changes the legal meaning of the same string.

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

One diagnostic ablation disables the discrepancy-posting validators that normally allow a later check to trigger targeted re-evaluation. This reduces the reference system to a single forward pass over the same Knowledge Sources. On the relevant diagnostic cases, ambiguous personal-expense resolution fell from 100% to 82%. The ablation is not a separate public benchmark split, so it should be read as system analysis rather than as a leaderboard result. It is nevertheless useful because it shows that the gain is not merely from decomposing the task into smaller modules. The ability to post a contradiction, wake a specialised resolver, and supersede an earlier hypothesis is part of the neurosymbolic property being tested.

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 appear competent at a row or category level and still fail every exact ledger case.

The results support three claims. First, symbolic spreadsheet evidence should be preserved as first-class data rather than collapsed into text. Second, LLMs are valuable when used as bounded semantic heuristics over typed uncertainty. Third, strict all-box scoring is necessary for regulated document tasks because partial correctness can conceal deployment-blocking errors.

Security and Prompt Injection

The benchmark includes adversarial-looking spreadsheet content but is not a full security benchmark. The reference architecture has useful natural defences: symbolic sources fire before LLM sources; LLMs see only ambiguous residual rows; LLM responses are schema-validated against closed category enumerations; amounts are not accepted from the model as new source values; and validators can post discrepancies when totals cease to balance. These guards should be treated as requirements for any LLM-assisted system evaluated on this task. A separate prompt-injection benchmark for spreadsheet-carried instructions would be valuable future work.

Why Priority Scheduling Is Used

Recent LLM-blackboard systems often use volunteer routing, where agents decide whether to act on a posted request. The reference system instead uses deterministic priorities. This is a design trade-off. Volunteer routing may improve open-world flexibility; deterministic priorities improve reproducibility, auditability, and regression testing. The benchmark can support either approach, but the scorer rewards final evidence-grounded correctness rather than a particular scheduler.

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. This is why the benchmark package exposes the scorer and fixture generator. It is designed to make future failures informative rather than to protect the current reference system from competition.

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 token counts were recorded during the original run; blackboard token counts were reconstructed on a later re-run over the same held-out

cases. The token ratio should therefore be read as approximate economic evidence, not as a single perfectly controlled timing experiment.

Artefact release. The benchmark is designed for open release, but any submission should include a stable repository URL, version number, licence, and clear data statement. If anonymised review is required, an anonymised review package should be provided first and linked to a permanent archive after acceptance.

Reproducibility Package

The submission package should make the benchmark executable without private TaxBridge infrastructure. The minimum package contains:

- generated workbooks for each corpus and the corresponding source `.spec.json` files where applicable;
- ground-truth target JSON with row-level truth, MTD box totals, expected flags, and workbook-shape metadata;
- scorer code with the pence tolerance, dynamic denominators, fungible-box equivalence rules, and row-level release checks;
- baseline prompts, serialised-workbook inputs, raw model outputs, model identifiers, token counts, and per-case scorer reports;
- a manifest stating version number, generation date, licence, supported tax year, and intended train/test usage.

The artefacts should also include failure cases. The purpose of the benchmark is not to curate examples that the current system already handles cleanly, but to preserve realistic accounting edge cases that expose defects in future systems. For the same reason, generated workbooks should remain inspectable as ordinary Excel files rather than only as JSON fixtures. Reviewers should be able to open a workbook, inspect the visual layout, and trace a reported total back to the row-level truth.

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 (1986). Truth-maintenance systems provide the background for defeasible revision and dependency tracking Doyle (1979); de Kleer (1986). Recent LLM blackboard work revives the pattern for agent coordination, often keeping LLMs as the primary reasoners Salemi et al. (2025); Han and Zhang (2025). 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 Tian et al. (2024); Chen et al. (2025); Xia et al. (2024). Their emphasis is representation and general spreadsheet reasoning. 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 is a neurosymbolic benchmark for spreadsheet-to-ledger reasoning under regulatory constraints. Its main lesson is that exact financial document reconstruction depends on evidence that text-only LLM serialisation tends to discard. The benchmark therefore provides a concrete testbed for systems that combine symbolic spreadsheet parsing, bounded neural semantic judgement, defeasible correction, and auditable aggregation. The current reference system solves the blind split and deployment corpus, while LLM-first baselines obtain partial but not submission-ready results. Future work should add stronger tool-using agent baselines, a dedicated prompt-injection track, and independently contributed systems.

Data Availability

The benchmark package should be submitted with a stable repository URL before journal submission. The intended package contains generated workbooks, `.spec.json` fixtures, ground-truth JSON, scorer code, baseline prompts, raw baseline outputs, and summary reports. An anonymised review URL or DOI should be inserted here: **[ANONYMISED/STABLE BENCHMARK URL TO INSERT BEFORE SUBMISSION]**.

References

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.

Nii, H. P. (1986). Blackboard systems: The blackboard model of problem solving and the evolution of blackboard architectures. *AI Magazine*, 7(2), 38–53; 7(3), 82–106.

Doyle, J. (1979). A truth maintenance system. *Artificial Intelligence*, 12(3), 231–272.

de Kleer, J. (1986). An assumption-based TMS. *Artificial Intelligence*, 28(2), 127–162.

Salemi, A., Parmar, M., Goyal, P., Song, Y., Yoon, J., Zamani, H., Palangi, H., & Pfister, T. (2025). LLM-based multi-agent blackboard system for information discovery in data science. *arXiv preprint arXiv:2510.01285*.

Han, B., & Zhang, S. (2025). Exploring advanced LLM multi-agent systems based on blackboard architecture. *arXiv preprint arXiv:2507.01701*.

Tian, Y., Zhao, J., Dong, H., Xiong, J., Xia, S., Zhou, M., Lin, Y., Cambronero, J., He, Y., Han, S., & Zhang, D. (2024). SpreadsheetLLM: Encoding spreadsheets for large language models. *arXiv preprint arXiv:2407.09025*.

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 Web Conference (WWW) 2025*. arXiv:2403.03636.

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 *Proceedings of the 3rd Workshop on Advances in Language and Vision Research*. arXiv:2405.16234.