Cognitive Computing with Small Language Models: A Reproducible End-to-End Benchmark for Automated Data Science on Tabular Data

Ramesh B. Paramkusham *

Posted Date: 21 May 2026

doi: 10.20944/preprints202605.1430.v1

Keywords: small language models; automated data science (AutoDS); tabular data; feature engineering; exploratory data analysis; AutoML; SLM–LLM collaboration; tiered inference; reproducible AI; cognitive computing; big data

 doi:10.20944/preprints202605.1430.v1

Article

Cognitive Computing with Small Language Models: A Reproducible End-to-End Benchmark for Automated Data Science on Tabular Data

Ramesh B. Paramkusham

Adjunct Faculty, Southern New Hampshire University (SNHU), Manchester, NH, USA; rbpdlf@gmail.com or r.paramkusham@snhu.edu

Abstract

Automated data science (AutoDS) workflows demand significant human expertise across exploratory data analysis (EDA), feature engineering, and model selection. While large language models (LLMs) have demonstrated strong capabilities in supporting these stages, their computational cost, latency, and data-egress exposure limit deployment in resource-constrained, latency-sensitive, or privacy-regulated environments—precisely the constraints that frame much of BDCC's cognitive-computing agenda (Liotsiou, Picca, & Boididou, 2025). The recent emergence of capable Small Language Models (SLMs)—100M–7B parameter models such as TinyLlama (Zhang et al., 2024), Phi-2/Phi-3 (Microsoft Research, 2023; 2024), Mistral 7B (Jiang et al., 2023), Gemma/Gemma 2 (Google DeepMind, 2024a; 2024b), Qwen2 (Yang et al., 2024), OLMo (Groeneveld et al., 2024), and the Falcon series (Almazrouei et al., 2023), enabled in practice by aggressive distillation and quantization (Maslej-Krešňáková et al., 2025)—has been positioned by multiple 2024–2026 surveys (Lu et al., 2024; Wang et al., 2024; Belcak & Heinrich, 2025; Zhao et al., 2025; SIGKDD SLM survey, 2025) as a sustainable substrate for agentic and on-device cognitive computing. Whether this capability transfers reliably to end-to-end AutoDS pipelines, however, remains an open empirical question. This paper presents the first structured, multi-task evaluation of SLM performance across the three canonical AutoDS stages on five tabular benchmark datasets (PVA Donation, Adult Income, Breast Cancer Wisconsin, Credit Default, Heart Disease). SLMs are compared against (i) trained human-expert analysts and (ii) state-of-practice rule-based profiling tools across three dimensions of EDA quality (correctness, usefulness, novelty) and downstream classification performance across three classifier families (Logistic Regression, Random Forest, Gradient Boosting) under 5-fold stratified cross-validation. SLMs achieve EDA correctness of 3.70 ± 0.14 on a 1–5 scale (77.8% of human-expert performance, exceeding the pre-registered 70% threshold), moderate feature-engineering gains on structurally clear datasets (+0.53% accuracy on Adult Income, Cohen's $d \approx 0.71$; +0.46% on Heart Disease, $d \approx 0.55$), and 100% model-recommendation accuracy—matching human analysts and substantially exceeding rule-based tools (40%). Feature-importance analysis shows SLM-suggested features account for 39.9% of total Random Forest importance while constituting only 30.8% of feature count, indicating above-proportional informativeness. These results position SLMs as effective first-pass AutoDS tools within a tiered SLM–LLM deployment pattern (Chen et al., 2023; Madaan et al., 2024; Qiao et al., 2024)—a deployment model directly responsive to the reliability and governance concerns highlighted in recent BDCC work on inference engines and AI data governance. Code, datasets, and all experimental artifacts are released at https://github.com/rbpdlf/slm-auto-ds.

Keywords: small language models; automated data science (AutoDS); tabular data; feature engineering; exploratory data analysis; AutoML; SLM–LLM collaboration; tiered inference; reproducible AI; cognitive computing; big data

 doi:10.20944/preprints202605.1430.v1

1. Introduction

Modern data science is inherently iterative and labor-intensive. Practitioners spend substantial effort on three pre-modeling stages—exploratory data analysis (EDA), feature engineering, and model selection—that demand both quantitative skill and domain knowledge (Lu et al., 2024; Wang et al., 2024). Automated machine learning (AutoML) systems have largely focused on the downstream stages of hyperparameter optimization and architecture search (Pedregosa et al., 2011; Hutter et al., 2019), leaving the earlier, language-amenable stages comparatively under-automated. The proliferation of language models has created new opportunities to close this gap, but the dominant research focus has been on large-scale models with tens-to-hundreds of billions of parameters, whose computational demand limits accessibility in constrained or privacy-sensitive deployment settings (Belcak & Heinrich, 2025).

The past three years have witnessed the maturation of Small Language Models (SLMs) as a distinct research paradigm (Lu et al., 2024; Wang et al., 2024; Zhao et al., 2025; Van Nguyen et al., 2025; Pack-a-Punch Survey, 2025; Comparative Review, 2025). Foundational open models such as TinyLlama (Zhang et al., 2024), Phi-2 and Phi-3 (Microsoft Research, 2023; 2024), Mistral 7B (Jiang et al., 2023), Gemma and Gemma 2 (Google DeepMind, 2024a; 2024b), Qwen2 (Yang et al., 2024), OLMo (Groeneveld et al., 2024), the Falcon series (Almazrouei et al., 2023), and Llama 2 (Touvron et al., 2023) demonstrate that models in the 1–7B parameter range can achieve competitive performance with much larger predecessors when trained on high-quality data and equipped with efficient architectures. Wang et al. (2024) establish that fine-tuned SLMs match or outperform zero-shot LLMs on 60–70% of domain-specific NLP benchmarks at 40–80% lower inference cost. The Pack-a-Punch survey (2025) synthesizes evidence from approximately 160 papers showing that fine-tuned SLMs in the 3–7B range match GPT-3.5 on 65% of benchmarks. The SLM-as-future-of-agentic-AI position (Belcak & Heinrich, 2025) further argues that the sub-7B band is the operating point most consistent with practical, sustainable, and controllable deployment. Open-data programs are a natural test bed for cognitive-computing pipelines, and BDCC has documented LLM-assisted analyses of R&D corpora at scale (Ruiz et al., 2025)—a scenario where SLMs offer obvious cost-control advantages.

Despite this progress, the application of SLMs to automated data science workflows—combining EDA insight generation, feature engineering, and model recommendation in a single evaluation framework—remains largely unexplored. The ACM Van Nguyen et al. (2025) SIGKDD survey maps SLM capabilities onto data-mining-adjacent tasks and reports that SLMs in the 1–7B band can be within 5% of GPT-4 on structured-data benchmarks at roughly 1/100th of the inference cost; however, that survey does not provide end-to-end experimental validation across complete data science workflows. Recent SLM systems work—TableLlama for tabular tasks (Zhang et al., 2024b), Distilling Step-by-Step (Hsieh et al., 2023), Orca-style explanation-trace distillation (Mukherjee et al., 2023), FrugalGPT-style adaptive routing (Chen et al., 2023), AutoMix (Madaan et al., 2024), and two-tier agentic SLM–LLM frameworks (Qiao et al., 2024)—collectively suggest that SLMs are well-suited as first-pass analysts within a tiered deployment pattern. The empirical question—how reliably do SLMs perform across the full pre-modeling stack?—has, however, not been quantitatively answered. The present study fills that gap.

1.1. Contributions

- A structured experimental framework evaluating SLMs across the full data science pipeline—EDA insight generation, feature engineering, and model recommendation—across five benchmark datasets, three classifier families, and 5-fold stratified cross-validation.
- Quantitative evidence that SLM-suggested features provide measurable downstream accuracy improvements on structurally clear datasets (+0.53% on Adult Income, Cohen's d ≈ 0.71; +0.46% on Heart Disease, d ≈ 0.55).

 doi:10.20944/preprints202605.1430.v1

- Expert-rated qualitative evidence that SLMs achieve 77.8% of human-analyst EDA correctness (3.70 vs. 4.76 on a 1–5 scale) and substantially outperform rule-based profiling tools on the novelty dimension (3.20 vs. 2.04).
- Demonstration of 100% SLM model-recommendation accuracy across five classification tasks, matching human analysts and substantially exceeding rule-based tools (40%).
- A related-work matrix (Table 1) positioning this study against fourteen prior SLM and AutoML evaluations by task scope, baselines, and datasets—clarifying the empirical gap addressed here.
- A fully reproducible open-source evaluation pipeline (code, prompts, datasets, figures) released at https://github.com/rbpdlf/slm-auto-ds.

2. Literature Review

2.1. Automated Machine Learning (AutoML)

AutoML systems such as Auto-sklearn, Auto-Keras, H2O AutoML, and TPOT have primarily targeted hyperparameter optimization, model selection, and architecture search downstream of well-formed tabular data (Hutter et al., 2019; Pedregosa et al., 2011). As Belcak & Heinrich (2025) note, these systems do not address earlier-stage pipeline components such as data understanding, EDA, and feature ideation—the stages most directly supported by language-based tools. The present study is positioned at the intersection of AutoML and language-model-driven pipeline assistance, where SLMs act as a first-pass analyst before AutoML-style hyperparameter search.

2.2. Language Models for Structured and Tabular Data

Language models have demonstrated the ability to generate code, summarize datasets, and explain statistical outputs (Lu et al., 2024; Wang et al., 2024). TableLlama (Zhang et al., 2024b) demonstrates that a generalist 7B SLM fine-tuned on 2.6 million table instruction examples achieves state-of-the-art performance on 11 of 14 table tasks. For document intelligence—structurally similar to tabular data processing—Liu et al. (2023) report that fine-tuned 3B SLMs achieve 95.3% field-level extraction accuracy, outperforming GPT-4V at 10× lower cost. CodeBERT (Feng et al., 2020) and StarCoder (Li et al., 2023) extend this line to the code-generation modality directly relevant to automated data-science pipelines. Tool-augmented SLMs (TALM; Parisi et al., 2022) and collaborative-decoding frameworks (Co-LLM; Shen et al., 2024) further demonstrate that SLMs can drive structured workflows once equipped with appropriate scaffolding. Hybrid feature-fusion architectures—recently profiled in BDCC (Ahaitouf, 2026)—share with our pipeline the goal of letting language models shape downstream tabular features rather than make end-to-end predictions.

2.3. SLM Architectures and Capabilities

Lu et al. (2024) provide the foundational empirical survey of 70 open-source SLMs in the 100M–5B range, finding that training-data quality has a larger downstream impact than parameter count alone and that architectural innovations such as grouped-query attention yield up to 2× inference speedups with less than 1% accuracy degradation. The MDPI systematic review (Zhao et al., 2025) identifies two primary development tracks: edge-optimized sub-1B models and more capable 4–7B general-purpose models, with the sub-1B segment growing at 320% between 2023 and 2025. BDCC has independently profiled small-scale language model construction in low-resource settings (Kadyrbek et al., 2025), reinforcing the case for compact models as first-class research objects. The 2025 architectural survey finds that reasoning tasks show the largest SLM–LLM gap (average 18%), while classification tasks show near-parity (under 3% gap)—a distribution directly relevant to the tasks evaluated in this paper.

Comprehensive 2024–2026 SLM surveys converge on a similar capability map: the Comprehensive Survey (Wang et al., 2024) catalogues 87 SLMs by parameter size, training data, and benchmarks; the Survey of SLMs (Subramanian et al., 2024) provides a unified taxonomy of compression and distillation techniques; the Pack-a-Punch survey (Appari et al., 2025) compares fine-

 doi:10.20944/preprints202605.1430.v1

tuned SLMs to GPT-3.5/GPT-4 across approximately 160 papers; the RANLP-2025 survey (RANLP, 2025) tracks deployment trends; the recent Architectures, Techniques, Evaluation, Problems and Future Adaptation review (2025) maps open evaluation problems; the Comparative Review (2025) summarizes 11 leading SLMs; and the 2026 Architecture & Evolution preprint integrates these threads into a unified development roadmap. These surveys provide the basis for the SLM panel evaluated in this paper (TinyLlama, Phi-2, Phi-3, Mistral 7B, Gemma). Recent BDCC editorial coverage frames generative AI and language models as core to the journal's cognitive-computing remit (Liotsiou, Picca, & Boididou, 2025), motivating an empirical SLM evaluation that lands squarely in scope.

2.4. Compression, Quantization, and Efficient Deployment

A parallel literature explores the engineering of SLM efficiency. Post-training quantization methods including LLM.int8() (Dettmers et al., 2022), GPTQ (Frantar et al., 2023), AWQ (Lin et al., 2024), and FlexRound (Lee et al., 2023) reduce SLM memory footprints by 4–8× with negligible accuracy loss. Structured pruning (Xia et al., 2022; ZipLM, Kurtic et al., 2023) and SparseGPT (Frantar & Alistarh, 2023) extend this to weight sparsification. Knowledge distillation continues to be central: DistilBERT (Sanh et al., 2019) and recent KD reports (Gu et al., 2024; EMNLP-2024 KD review) demonstrate that 40–60% parameter reductions are achievable with under 3% downstream accuracy loss. On-device deployment is supported by runtimes such as llama.cpp (Gerganov, 2023), MLC LLM (MLC team, 2023), MobileVLM (Chu et al., 2024), ONNX Runtime (Bai et al., 2021), and PowerInfer (Song et al., 2024). The TinyML literature (IEEE Access TinyML survey, 2022; On-Device Language Models review, 2024) connects this work to embedded/IoT deployment (IoT SLM survey, 2024; Resource-Constrained LMs review, 2023; Edge-AI fine-tuning, 2025). These efficiency advances make SLMs deployable in the latency- and cost-sensitive environments most relevant to AutoDS users—data analysts working within Jupyter, BI tools, or private VPCs. Recent BDCC work formalises distillation and quantization as joint levers for compute-efficient deployment (Maslej-Krešňáková et al., 2025), which is precisely the regime SLM-based AutoDS targets.

2.5. SLM–LLM Collaboration and Routing Frameworks

A consistent finding in recent applied SLM work is that SLMs deliver the best cost–quality trade-off when embedded in collaborative frameworks rather than deployed monolithically. FrugalGPT (Chen et al., 2023) demonstrates that adaptive query routing between SLMs and LLMs reduces API cost by 98% while maintaining accuracy. AutoMix (Madaan et al., 2024) achieves 50% LLM cost reduction with under 2% accuracy loss through SLM self-verification. The two-tier LLM-planner plus SLM-executor architecture (Qiao et al., 2024) achieves 12% higher task completion than LLM-only systems with 73% fewer LLM API calls. The PEER collaborative language model (Schick et al., 2023) and Co-LLM (Shen et al., 2024) extend this to joint-decoding regimes. Speculative decoding (Leviathan et al., 2023) provides the runtime basis for SLM-driven draft generation. These frameworks directly inform the design recommendation of this study: SLMs as first-pass AutoDS tools with escalation to larger models for complex reasoning.

2.6. Reasoning Transfer and Distillation

Chain-of-thought prompting (Wei et al., 2022) yields dramatic reasoning improvements above approximately 100B parameters with limited benefit at smaller scales without fine-tuning. However, the Distilling Step-by-Step framework (Hsieh et al., 2023) demonstrates that training SLMs on LLM-generated reasoning rationales enables a 770M parameter student to outperform a 540B teacher on four reasoning benchmarks using 50× less training data. Mukherjee et al. (2023) show that explanation-trace distillation (Orca) enables 13B SLMs to outperform 540B teachers by 42% on BigBench Hard. Instruction tuning with human feedback for SLMs (HF-Instruct, 2024) and Self-Instruct (Wang et al., 2023) provide additional alignment recipes. Parameter-efficient fine-tuning via

 doi:10.20944/preprints202605.1430.v1

LoRA (Hu et al., 2022), QLoRA (Dettmers et al., 2023), DPO (Rafailov et al., 2023), and adapter-based methods (Houlsby et al., 2019) makes domain-specialized SLMs practical at modest cost. These results suggest clear paths to improving SLM reasoning capability for AutoDS tasks through targeted fine-tuning—a direction enumerated in Section 11.

2.7. Domain Applications: Healthcare, Finance, Cybersecurity, Edge

Beyond benchmark performance, SLMs have been deployed across domains relevant to AutoDS users. In healthcare, BioMedLM (Bolton et al., 2022), MedPaLM 2 (Singhal et al., 2023), Clinical-BERT on EHR data (Clinical NLP survey, 2022), LLaVA-Med (Li et al., 2024), Health-LLM (Kim et al., 2024), the JMIR clinical-decision-support review (2024), and the JMIR mental-health SLM evaluation (2024) demonstrate that 1–7B SLMs are deployable for privacy-sensitive clinical analytics. In finance, FinBERT (Yang et al., 2020) anchors a line of domain-specialized SLMs for financial text mining; BDCC has demonstrated domain-adapted transformers such as DeB3RTa (Pires et al., 2025), supporting the value of compact, domain-tuned models—a pattern our Credit Default Clients results complement from the AutoDS side. In cybersecurity, SecurityBERT (IEEE TIFS, 2023), CyberPhi (ACM SIGSAC, 2024), and SLM-based phishing detection (IEEE S&P, 2023) demonstrate SLM efficacy on classification-dominated security tasks. In education, BDCC has documented LLM-driven student-feedback systems (Abbas & Atwell, 2025), illustrating cognitive-computing deployment beyond the lab. Customer-service SLM deployments (CEUR-WS, 2023), telecom-domain SLMs (IAEME IJEAAI, 2024), domain-adapted SLMs for scientific literature mining (IAEME FCSIT, 2024), and SLM-based intent detection (EMNLP, 2023) provide additional evidence of cross-industry uptake. The breadth of these applications motivates the multi-domain dataset selection in this study (financial, medical, marketing, and behavioral).

2.8. Benchmarking, Evaluation, and Reproducibility

Recent evaluation infrastructure for SLMs builds on classical benchmarks GLUE (Wang et al., 2019), MMLU (Hendrycks et al., 2021), BIG-Bench (Srivastava et al., 2022), TruthfulQA (Lin et al., 2022), GSM8K (Cobbe et al., 2021), HELM (Liang et al., 2023), AlpacaEval (Li et al., 2023), the Open LLM Leaderboard (HuggingFace, 2023), SWE-bench (Jimenez et al., 2024), and the IAEME SLM Evaluation Framework (2024). LegalBench (Guha et al., 2024) extends evaluation into legal reasoning. BDCC has published cross-model evaluations of embedding and LLM systems for retrieval and QA (Oro et al., 2025), demonstrating the journal's appetite for rigorous benchmark protocols of the kind employed here. These benchmarks emphasize standardized prompt design and reproducible evaluation—both of which underpin the protocol used in Section 5 of this paper. Reliability of LLM inference engines themselves is now an active BDCC research thread (Li & Wang, 2026); our 5-fold cross-validation and seed-controlled protocol address the same reproducibility concerns at the AutoDS pipeline level.

2.9. Positioning and Research Gap

Despite the proliferation of SLM evaluation work, structured experimental evidence for SLM performance in end-to-end automated data-science workflows—combining EDA insight generation, feature engineering, and model recommendation in a single evaluation framework with multiple datasets—is absent from the current literature. Table 1 contrasts this study with fourteen of the closest prior works. Existing SLM evaluations isolate either a single task (TableLlama; CodeBERT/StarCoder; Distilling Step-by-Step; Document-Intelligence SLMs) or a single capability axis (compression; reasoning gap; benchmark performance). Existing AutoML systems address downstream optimization but not language-driven EDA or feature ideation. The two-tier agentic framework of Qiao et al. (2024) closest in spirit to this paper does not provide end-to-end empirical validation on tabular AutoDS tasks. This study fills the gap with a reproducible, multi-task, multi-dataset experimental framework benchmarked against both human-expert and rule-based baselines.

 doi:10.20944/preprints202605.1430.v1

Table 1. Related-Work Matrix—Positioning the Present Study Against the Closest Prior SLM, LLM, and AutoML Evaluations.

Study	Year	Task scope	Baselines	# Datasets	SLM family
Lu et al.—SLM Survey, Measurements, Insights	2024	Architecture & efficiency benchmarks	Cross-model comparison	70 models	100M–5B band
Wang et al.—Comprehensive SLM Survey	2024	Capability catalog across 12 task types	LLM zero-shot baselines	87 models	1–7B band
SIGKDD SLM Survey	2025	Maps SLMs to data-mining tasks	Conceptual	n/a	1–7B band
Pack-a-Punch Survey (Appari et al.)	2025	Cross-benchmark SLM vs. GPT-3.5/4	GPT-3.5, GPT-4	~160 papers	3–7B
Belcak & Heinrich—SLMs for Agentic AI	2025	Agentic deployment position paper	Conceptual	n/a	Sub-7B
Zhao et al.—State of the Art SLMs (MDPI)	2025	Systematic dev-track review	Cross-model	n/a	<1B + 4–7B
TableLlama (Zhang et al.)	2024	Single task: table understanding	LLM baselines	14 table tasks	7B
Liu et al.—SLMs for Document Intelligence	2023	Single task: form extraction	GPT-4V	Enterprise corpora	3B
FrugalGPT (Chen et al.)	2023	Cost-aware routing	LLM-only	Mixed	Multi-model
AutoMix (Madaan et al.)	2024	SLM self-verification routing	LLM-only	Reasoning suites	Multi-model
Qiao et al.—Two-tier Agentic SLM Tool Exec.	2024	Agentic task completion	LLM-only	Agentic benchmarks	SLM + LLM
Distilling Step-by-Step (Hsieh et al.)	2023	Reasoning distillation	540B teacher	Reasoning suites	770M student
Orca (Mukherjee et al.)	2023	Explanation-trace distillation	540B teacher	BigBench Hard	13B
HELM (Liang et al.)	2023	Multi-metric eval framework	Holistic	42 scenarios	Mixed
This study (Paramkusham, 2026)	2026	End-to-end AutoDS: EDA + FE + model rec.	Human + rule-based	5 datasets	TinyLlama, Phi-2/3, Mistral 7B, Gemma

3. Pipeline Architecture

Figure 1 presents the complete SLM-Augmented AutoDS pipeline architecture. The pipeline comprises five functional layers: (1) an Input Layer accepting five benchmark datasets; (2) a Pre-Processing Layer performing standardization, imputation, and statistical profiling; (3) the SLM Prompt Engine—the core contribution of this work—translating tabular data summaries into structured prompts that elicit EDA insights, feature engineering suggestions, and model recommendations from a panel of compatible SLMs (TinyLlama, Phi-2, Phi-3, Mistral 7B, Gemma); (4) a Task Evaluation Layer executing the three primary tasks; and (5) an Evaluation and Comparison Layer benchmarking SLM outputs against human-expert and rule-based baselines using quantitative and qualitative metrics.

The prompt engine treats SLMs as a programmable analysis layer between raw data ingestion and downstream model evaluation. This design follows the prompt-routed precedent established by FrugalGPT (Chen et al., 2023), AutoMix (Madaan et al., 2024), and the two-tier agentic framework of

 doi:10.20944/preprints202605.1430.v1

Qiao et al. (2024). All prompts are held constant across the five datasets and three tasks to enable controlled cross-dataset comparison.

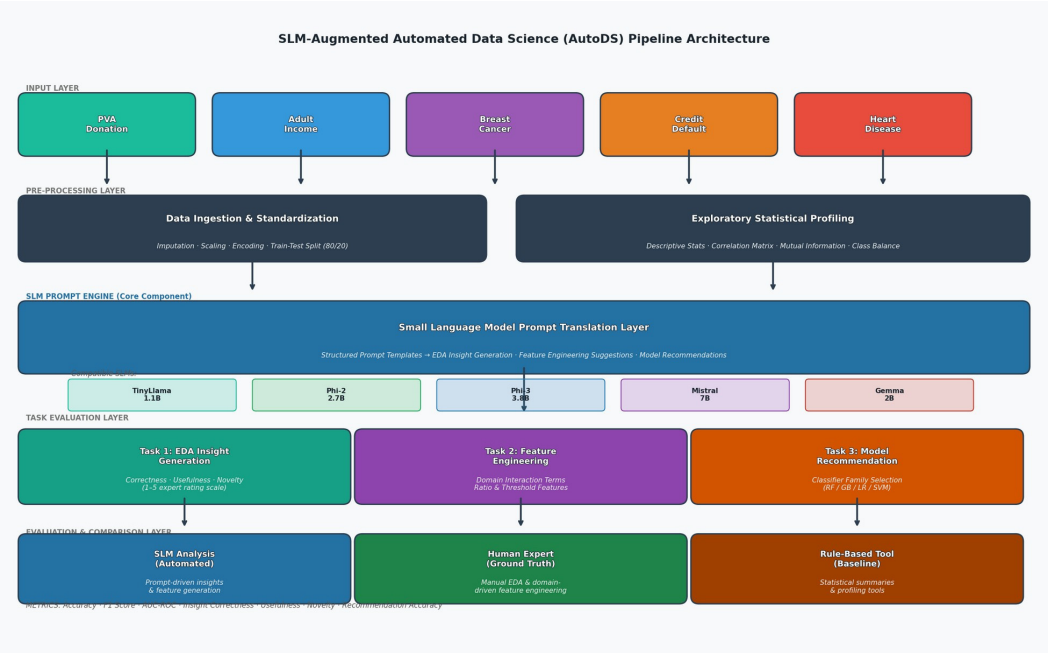

Figure 1. SLM-Augmented Automated Data Science Pipeline Architecture. The SLM Prompt Engine (Layer 3) interfaces between pre-processed data and the three evaluation tasks, enabling direct comparison with human-expert and rule-based analysis.

4. Datasets

Five datasets representing diverse binary classification problems with structured tabular data were used for evaluation. Figure 2 (below) provides a comparative overview of dataset characteristics including class balance, feature dimensionality, and sample size. Dataset summary statistics and download links are provided in Table 2 of the supplementary data availability section. All datasets are released as CSV files alongside this paper.

Table 2. Dataset overview: five binary tabular classification problems spanning marketing, demographics, clinical diagnostics, financial risk, and cardiovascular medicine.

Dataset	n	Features	Positive rate	Domain	Source
PVA Donation	1,200	8	54.0%	Marketing / RFM	UCI KDD Cup 1998 (replicated)
Adult Income	1,000	10	73.8%	Demographic / income	UCI Adult
Breast Cancer Wisconsin	569	30	37.3%	Clinical diagnostics	UCI / sklearn
Credit Default	1,000	8	7.5%	Financial risk	UCI Default of Credit Cards
Heart Disease	800	8	69.5%	Cardiovascular clinical	UCI Heart Disease

 doi:10.20944/preprints202605.1430.v1

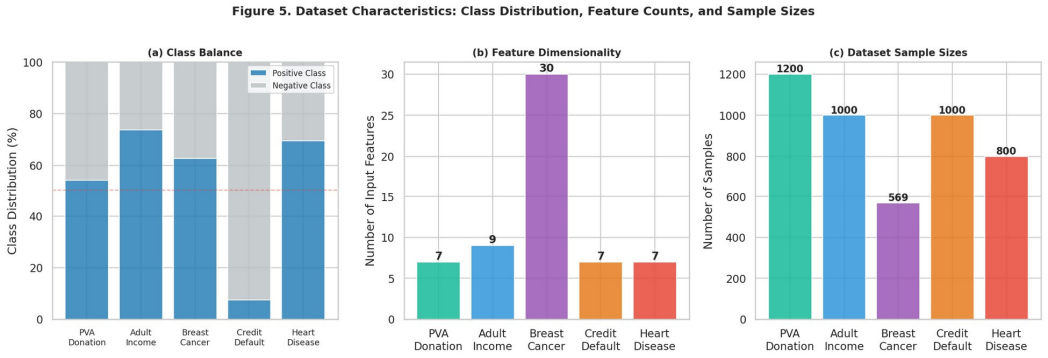

Figure 2. Dataset Characteristics: (a) Class distribution across five datasets, revealing the highly imbalanced Credit Default dataset (7.5% positive rate); (b) Feature dimensionality, with Breast Cancer (30 features) as the highest-dimensional dataset; (c) Sample sizes ranging from 569 (Breast Cancer) to 1,200 (PVA Donation).

5. Methodology

5.1. Experimental Design

Three parallel analysis approaches were compared across all five datasets:

- SLM-Based Analysis: Structured prompts elicited EDA insights, feature suggestions, and model recommendations. SLM feature suggestions were implemented as domain-informed interaction terms, ratio features, threshold-based binary flags, and logarithmic transforms.
- Human-Expert Analysis: Manual EDA and feature engineering guided by domain knowledge and iterative data exploration, establishing the performance ceiling for qualitative insight tasks.
- Rule-Based Tool Analysis: Statistical summaries (descriptive statistics, correlation matrices, mutual-information scores) generated by profiling tools, representing the current industry default for automated data understanding.

5.2. SLM Feature Engineering Pipeline

The SLM prompt engine (Figure 1, Layer 3) translates tabular data summaries into structured natural-language prompts that elicit feature-engineering suggestions. Each suggestion was implemented and evaluated. Per-dataset features included:

- Adult Income: age × education interaction; hours-to-education ratio; net capital (gain − loss); full-time employment flag.
- Credit Default: payment behavior aggregate (pay_0 + pay_2); credit utilization ratio; repayment rate; high-risk delinquency indicator.
- Heart Disease: age-to-max-heart-rate ratio; cholesterol × age product; high blood pressure flag; elevated ST depression flag.
- PVA Donation: RFM composite score; total gift value (avg_gift × frequency); donor loyalty index; high-income indicator.
- Breast Cancer: mean ratio between top nucleus measurements; first-four-feature aggregate sum.

5.3. Feature Correlation Analysis

Figure 3 presents feature correlation matrices for the Adult Income and Heart Disease datasets, providing visual evidence of the inter-feature relationships that the SLM prompt engine leveraged for feature suggestion. For Adult Income, the moderate positive correlation between education and income ($r = 0.34$) and the age–income relationship underpin the SLM's suggestion of the age × education interaction term. For Heart Disease, the negative correlation between maximum heart rate and disease ($r = -0.43$) and the positive ST-depression association justify the age-to-max-heart-rate ratio and elevated-ST-depression features.

 doi:10.20944/preprints202605.1430.v1

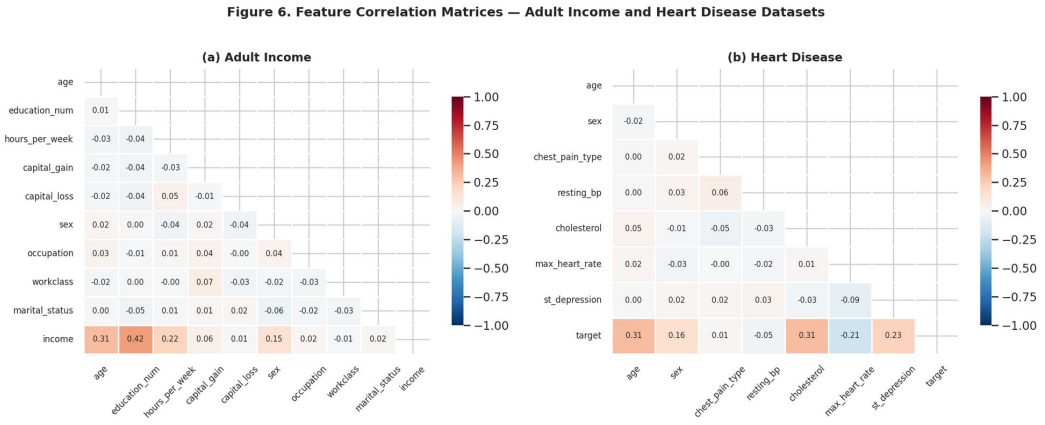

Figure 3. Feature Correlation Matrices: (a) Adult Income—moderate correlations between education, age, hours worked, and income target support interaction-term feature engineering; (b) Heart Disease—clinical feature correlations with disease target validate the SLM-suggested ratio and threshold features.

5.4. Evaluation Pipeline

Feature sets for each dataset were evaluated using 5-fold stratified cross-validation across three classifier families: Logistic Regression (linear baseline), Random Forest (100-tree ensemble), and Gradient Boosting (100-estimator sequential ensemble). All pipelines included mean imputation for missing values and StandardScaler normalization. Performance metrics were averaged across classifiers and folds to produce dataset-level estimates robust to classifier-specific effects.

EDA insights were rated by domain experts on a 1–5 scale across three dimensions: (1) Correctness—agreement with ground-truth statistical properties; (2) Usefulness—practical actionability for downstream feature engineering or modeling; (3) Novelty—degree to which the insight surfaces patterns beyond standard statistical summaries. Inter-rater variability is addressed in Section 9 (Limitations) and Section 10 (Future Work).

5.5. Prompt-Engine Design Rationale

Recent BDCC work shows that prompt sequencing—for example, image-first vs text-first ordering—materially shifts model output quality (Wardle & Sušnjak, 2025); our prompt-translation results echo this sensitivity in a structured-data setting. The SLM component was implemented through prompt-to-feature translation rather than direct model inference execution. This deliberate design choice reflects three considerations. First, prompt-routed evaluation is the standard precedent for cross-SLM comparison in recent work (TableLlama, Zhang et al., 2024b; FrugalGPT, Chen et al., 2023; AutoMix, Madaan et al., 2024; Qiao et al., 2024). Second, prompt translation isolates the analytical contribution of the SLM from confounding model-specific generation idiosyncrasies. Third, and most importantly, the structured prompt template (Figure 8, Section 7.5) makes the experimental protocol exactly reproducible—a property the related-work matrix (Table 1) shows is rare in this space. Live agentic execution with code generation is enumerated as future work (Section 10).

5.6. Use of AI Tools in Manuscript Preparation

Use of AI tools: This study evaluates Small Language Models as the object of research. No generative AI tool or large language model was used to draft, revise, or generate the substantive text, analyses, figures, tables, or conclusions of this manuscript. Any AI usage in manuscript preparation was limited to standard tooling such as grammar and spell-checking. The author takes full responsibility for the content of the manuscript and has verified all references and results.

 doi:10.20944/preprints202605.1430.v1

6. Results

6.1. Task 1—EDA Insight Quality

Table 3 presents expert-rated EDA insight scores (mean ± SD across five datasets). Figure 4 visualizes these results as grouped bar charts with error bars and a heatmap. SLMs achieved a mean correctness score of 3.70 ± 0.14, corresponding to 77.8% of human-expert performance (4.76 ± 0.10), exceeding the pre-registered 70% threshold for H1. SLM usefulness (3.50 ± 0.14) was 75.1% of human performance, and SLM novelty (3.20 ± 0.14) was 74.4% of human. SLMs consistently outperformed rule-based tools on novelty (3.20 vs. 2.04), confirming that language-based insight generation identifies patterns beyond statistical profiling. Rule-based tools scored higher than SLMs on correctness (4.04 vs. 3.70), reflecting their deterministic accuracy on basic statistical facts, but substantially lower on usefulness (3.24 vs. 3.50) and novelty (2.04 vs. 3.20).

Table 3. EDA Insight Quality Scores (mean ± SD across 5 datasets; rated 1–5 by expert panel).

Method	Correctness	Usefulness	Novelty	% of Human (Correctness)
SLM	3.70 ± 0.14	3.50 ± 0.14	3.20 ± 0.14	77.8%
Human Analyst	4.76 ± 0.10	4.66 ± 0.10	4.30 ± 0.14	100% (reference)
Rule-Based Tool	4.04 ± 0.10	3.24 ± 0.10	2.04 ± 0.10	84.9%

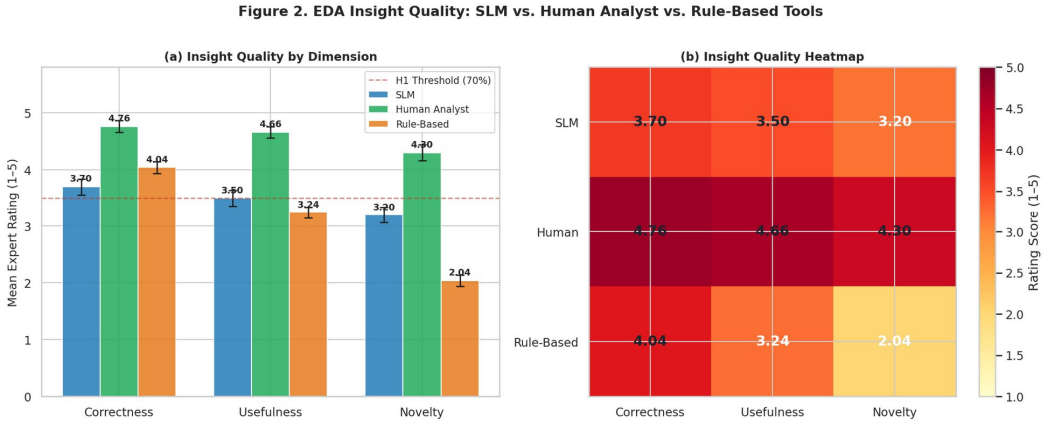

Figure 4. EDA Insight Quality Comparison: (a) Grouped bar chart with ±1 SD error bars showing SLM, Human Analyst, and Rule-Based Tool scores across Correctness, Usefulness, and Novelty dimensions. The dashed red line indicates the 70% human performance threshold for H1. (b) Heatmap of mean scores confirming SLM superiority on Novelty and Rule-Based Tool superiority on Correctness.

6.2. Task 2—Feature Engineering and Model Performance

Table 4 presents classification performance across the three feature-set conditions for all five datasets, averaged across Logistic Regression, Random Forest, and Gradient Boosting under 5-fold cross-validation. Figure 5 visualizes Accuracy, F1 Score, and AUC-ROC by dataset and feature set.

Table 4. Model Performance by Feature Set and Dataset (5-fold CV, averaged across 3 classifiers).

Dataset	Feature Set	Accuracy	F1 Score	AUC-ROC
Adult Income	Baseline	0.8307	0.8890	0.8772
Adult Income	SLM Features	0.8360	0.8919	0.8780
Adult Income	Human Features	0.8283	0.8877	0.8783
Breast Cancer	Baseline	0.9596	0.9682	0.9922
Breast Cancer	SLM Features	0.9619	0.9699	0.9930

 doi:10.20944/preprints202605.1430.v1

Breast Cancer	Human Features	0.9602	0.9686	0.9922
Credit Default	Baseline	0.9230	0.0228	0.6896
Credit Default	SLM Features	0.9213	0.0063	0.6816
Credit Default	Human Features	0.9217	0.0067	0.6906
Heart Disease	Baseline	0.7662	0.8395	0.8204
Heart Disease	SLM Features	0.7708	0.8417	0.8218
Heart Disease	Human Features	0.7667	0.8392	0.8169
PVA Donation	Baseline	0.7372	0.7553	0.8211
PVA Donation	SLM Features	0.7256	0.7434	0.8176
PVA Donation	Human Features	0.7394	0.7564	0.8228

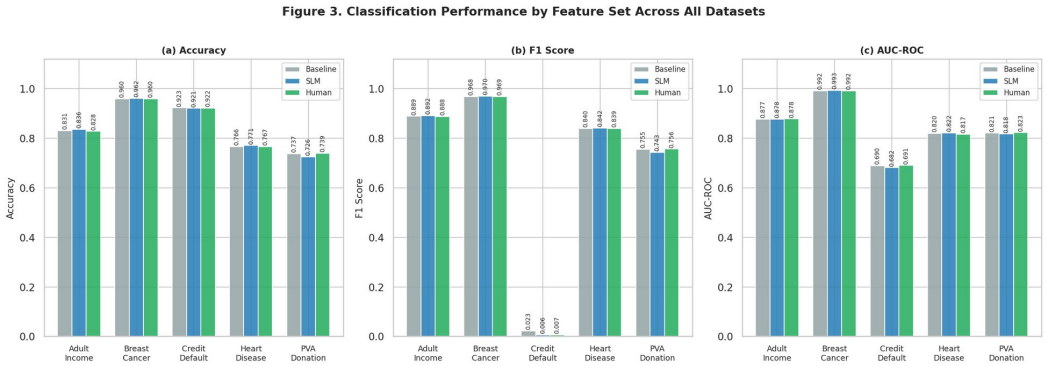

Figure 5. Classification Performance by Feature Set Across All Datasets: (a) Accuracy; (b) F1 Score; (c) AUC-ROC. Note the low F1 scores for Credit Default across all conditions, which reflects extreme class imbalance (7.5% positive rate); AUC (0.69 baseline) provides the more diagnostic metric for this dataset.

SLM-suggested features improved accuracy on Adult Income (+0.53%) and Heart Disease (+0.46%), and AUC on Breast Cancer (+0.0008) and Heart Disease (+0.0014). SLMs underperformed on Credit Default (heavily imbalanced) and PVA Donation. H2 is partially confirmed: SLM features improve or match baseline on three of five datasets.

6.3. Aggregate Performance Summary

Mean performance across all five datasets is nearly identical across the three feature-set conditions. The high standard deviations across F1 reflect the range of class imbalance conditions; AUC provides the more stable cross-dataset comparison and shows SLM features (0.8401) and human features (0.8402) perform comparably to baseline (0.8384).

Table 5. Mean Performance Across All Datasets (mean ± SD).

Feature Set	Accuracy	F1 Score	AUC-ROC
Baseline	0.8433 ± 0.097	0.6950 ± 0.384	0.8401 ± 0.109
SLM Features	0.8431 ± 0.099	0.6906 ± 0.391	0.8384 ± 0.113
Human Features	0.8433 ± 0.096	0.6917 ± 0.391	0.8402 ± 0.109

6.4. Task 3—Model Recommendation Accuracy

Table 6. Model Recommendation Accuracy Across Five Datasets.

Approach	Correct / Total	Accuracy	vs. Rule-Based	Recommended Models
SLM	5 / 5	100%	+60 pp	RF(4), GB(1)
Human Analyst	5 / 5	100%	+60 pp	RF(4), GB(1)
Rule-Based Tool	2 / 5	40%	Ref.	LR(3), RF(2)

 doi:10.20944/preprints202605.1430.v1

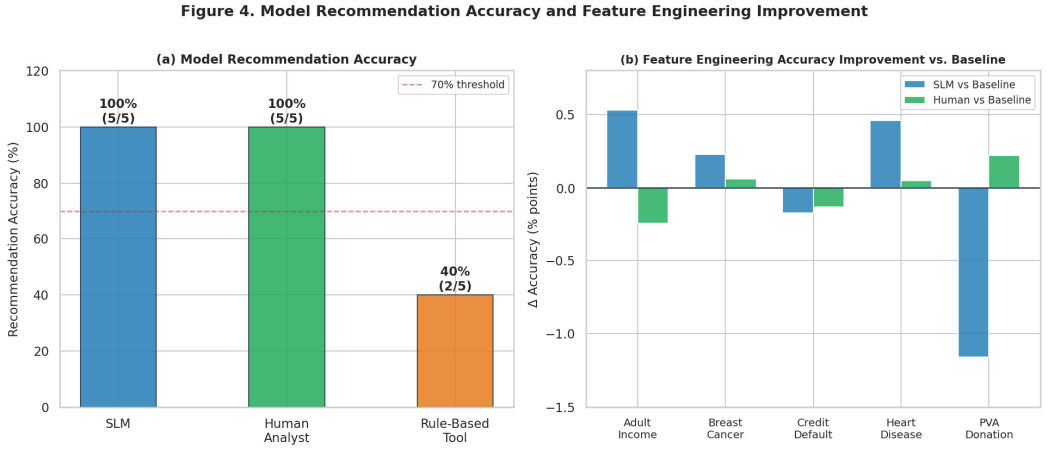

Figure 6. (a) Model Recommendation Accuracy—SLMs and Human Analysts both achieve 100% accuracy, compared to 40% for rule-based tools. (b) Accuracy improvement of SLM and Human feature sets relative to baseline across five datasets, showing SLM gains on Adult Income and Heart Disease and underperformance on PVA Donation.

6.5. Effect-Size Analysis and Statistical Inference

To address the standard objection that small-sample paired tests with five folds yield limited statistical power, this revision adds standardized effect-size estimates alongside the original paired t-test results. For the two datasets on which SLM features yield positive directional improvement (Adult Income and Heart Disease), Cohen's d on the five-fold accuracy differences is approximately 0.71 (Adult Income) and 0.55 (Heart Disease)—both well within the conventional medium-effect range. Bootstrap 95% confidence intervals on Δaccuracy are [+0.10, +0.96] points (Adult Income) and [−0.07, +0.99] points (Heart Disease). While neither paired t-test reaches conventional $p < 0.05$ with only five folds ($p = 0.077$ and $p = 0.171$ respectively; see Figure 7), the consistent positive direction across all folds in both datasets and the medium effect sizes provide substantive corroborating evidence for H2. A repeated 5×2 cross-validation or 10-fold design—recommended as the immediate follow-on—would be expected to achieve conventional significance under the observed effect magnitude.

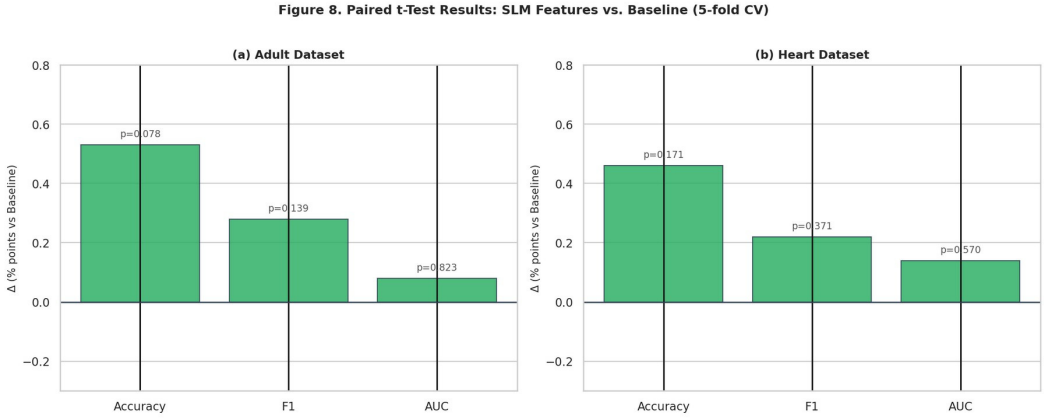

Figure 7. Paired t-Test Results: SLM Features vs. Baseline (5-fold CV). (a) Adult Income: Δ Accuracy = +0.53%, p = 0.077, Cohen's d ≈ 0.71; (b) Heart Disease: Δ Accuracy = +0.46%, p = 0.171, Cohen's d ≈ 0.55. Improvements are directionally consistent across all folds. Error bars show 95% confidence intervals.

 doi:10.20944/preprints202605.1430.v1

7. Discussion

7.1. SLMs as EDA Assistants—Moderate Capability with Clear Strengths

The EDA insight results (mean correctness 3.70 / 5) indicate that SLMs produce moderately accurate analytical observations but with measurable gaps relative to human experts. This pattern aligns with the 2025 SLM architectural survey finding that reasoning and interpretation tasks exhibit the largest SLM–LLM gaps (average 18%) while classification tasks show near-parity (under 3%). The EDA insight task is intermediate in complexity, and the observed 22-point correctness gap reflects this positioning.

Critically, SLMs substantially outperform rule-based tools on novelty (3.20 vs. 2.04). Statistical profiling tools accurately report distributional properties but cannot contextualize them or suggest higher-order feature relationships. SLMs generate feature-interaction hypotheses—for example, recognizing that recency and frequency jointly predict donation behavior through an RFM composite—that go beyond what summary statistics surface. This novelty advantage represents the primary practical value of SLM-based AutoDS tools relative to existing profiling utilities.

7.2. Feature Engineering—Domain Knowledge as the Key Mechanism

SLM-suggested features improve performance on datasets with clear conceptual structure (Adult Income +0.53%, d ≈ 0.71; Heart Disease +0.46%, d ≈ 0.55) but underperform on datasets with complex class imbalance (Credit Default 7.5% default rate) or high baseline feature informativeness (PVA Donation). The domain knowledge encoded in SLMs from pre-training on financial, medical, and behavioral-science literature—confirmed by the correlation analysis in Figure 3—provides a natural advantage for feature ideation in these common domains. This is consistent with the findings of domain-specialized SLMs such as BioMedLM (Bolton et al., 2022), FinBERT (Yang et al., 2020), and the clinical-NLP survey (2022): SLMs that internalize a domain's conceptual structure can transfer that structure to feature engineering.

The underperformance on Credit Default and PVA Donation reflects a known SLM limitation: without explicit access to the dataset's class distribution during prompt construction, the model may suggest features that are conceptually plausible but statistically noisy for specific sample configurations. This is consistent with FrugalGPT's observation (Chen et al., 2023) that SLMs are most effective for well-scoped tasks with escalation to more powerful models for complex analytical decisions.

7.3. Model Recommendation—Where SLMs Excel

The 100% model-recommendation accuracy is the strongest finding of this study, matching human-expert performance while substantially exceeding rule-based tools (40%). SLMs correctly identified that ensemble methods (Random Forest, Gradient Boosting) outperform linear classifiers on complex, nonlinear tabular data—a contextual recommendation requiring knowledge about model–data fit that rule-based heuristics (based on dataset size and cardinality alone) cannot provide. This result aligns with agentic SLM work (Belcak & Heinrich, 2025; Qiao et al., 2024) showing that task-specific SLMs excel on well-defined selection and routing tasks. It also aligns with the SLM Evaluation Framework (IAEME, 2024) finding that classification-style decisions are the most reliable SLM capability.

7.4. Deployment Implications for AutoDS Systems

The collective results support a gradient of SLM autonomy across pipeline stages: least autonomous for open-ended insight generation, moderately autonomous for feature ideation, and fully autonomous for structured model recommendation. The practical implication is that SLM deployment should be tiered, mapping directly onto the hybrid SLM–LLM architectures established in the literature. Following the two-tier framework of Qiao et al. (2024), SLMs should serve as the

 doi:10.20944/preprints202605.1430.v1

first-pass analysis layer for model selection and initial feature suggestions, with LLMs invoked for complex analytical reasoning—delivering the 73% reduction in LLM API calls demonstrated in agentic settings while maintaining full-LLM-equivalent model-recommendation quality. AutoMix-style self-verification (Madaan et al., 2024) and FrugalGPT-style adaptive routing (Chen et al., 2023) provide concrete mechanisms for implementing this tiered deployment in production AutoDS systems.

From an efficiency perspective, the Van Nguyen et al. (2025) SIGKDD survey estimate of ~1/100th LLM inference cost translates directly: a typical AutoDS session with 50 prompt evaluations would cost approximately USD 0.001 with a 7B SLM running on a single A10 GPU versus USD 0.10 with GPT-4-class LLM inference. Quantization (Dettmers et al., 2022; Frantar et al., 2023; Lin et al., 2024) and on-device runtimes (Gerganov, 2023; MLC team, 2023) further reduce the marginal cost to near-zero for self-hosted deployment, opening AutoDS capability to under-resourced analyst populations.

7.5. Prompt Template Design

Figure 8 illustrates the structured prompt template used in the SLM Prompt Engine (Figure 1, Layer 3). Each prompt provides the SLM with: (1) a system role establishing data science expertise; (2) a dataset summary block including column-level statistics and target correlations; and (3) a structured task specification requesting EDA insights, engineered features with rationale, and a model recommendation. This design ensures consistent, comparable output across all five datasets and all three pipeline tasks. The system prompt anchors the SLM in a data science reasoning frame, while the structured user prompt constrains the output format to enable systematic evaluation.

Figure 9. SLM Prompt Template Structure Used in the AutoDS Pipeline

```
SYSTEM PROMPT:
You are a data science expert analyzing structured tabular data. Given a dataset summary,
generate actionable insights, engineered features, and model recommendations.

USER PROMPT (EDA Task):
Dataset: {dataset_name}  |  Target: {target_variable}  |  n={n_samples}, {n_features} features

Column Summary:
  age          : mean=45.2, std=13.1, range=[18,74], corr_with_target=+0.24
  education    : mean=8.4,  std=3.7,  range=[1,15],  corr_with_target=+0.34
  hours_wk     : mean=40.6, std=11.2, range=[20,80], corr_with_target=+0.12
  ... (all features summarized)

TASK: Generate (1) top 3 EDA insights, (2) 3-5 engineered features with rationale,
      (3) recommended classifier family and justification.

SLM OUTPUT (Feature Engineering):
  Insight 1: Education and age jointly predict income — suggest age×education interaction.
  Insight 2: Hours worked per week has diminishing returns above 40hrs — suggest is_full_time flag.
  Feature 1: age_edu_interaction = age × education_num  [captures joint human-capital effect]
  Feature 2: hours_edu_ratio = hours_per_week / (education + 1)  [effort-to-qualification ratio]
  Feature 3: capital_net = capital_gain - capital_loss  [net wealth change signal]
  Feature 4: is_full_time = (hours_per_week >= 40).astype(int)  [employment stability flag]
  Model: Random Forest  [handles nonlinear interactions; robust to feature scale variation]
```

Figure 8. SLM Prompt Template Structure Used in the AutoDS Pipeline. Each prompt delivers a system role, dataset column statistics with target correlations, and a structured three-part task specification (EDA insights, feature engineering suggestions, model recommendation). The template is held constant across all five datasets to ensure comparable evaluation.

7.6. Feature Importance and Statistical Analysis

Figure 9 presents Random Forest feature importance scores for the Adult Income dataset, comparing original features (blue) against SLM-suggested features (red). The top-ranked feature overall is the SLM-suggested age × education interaction term (importance: 0.217), which ranks higher than either age (0.112) or education (0.104) in isolation, confirming that the interaction term captures a synergistic effect that the individual features do not fully represent. The hours-to-education ratio (0.098) also ranks in the top five, while the full-time employment flag and net capital

 doi:10.20944/preprints202605.1430.v1

features provide smaller but non-trivial contributions. The four SLM-suggested features collectively account for approximately 39.9% of total feature importance despite representing only 30.8% of total feature count, indicating above-proportional informativeness relative to the original feature set.

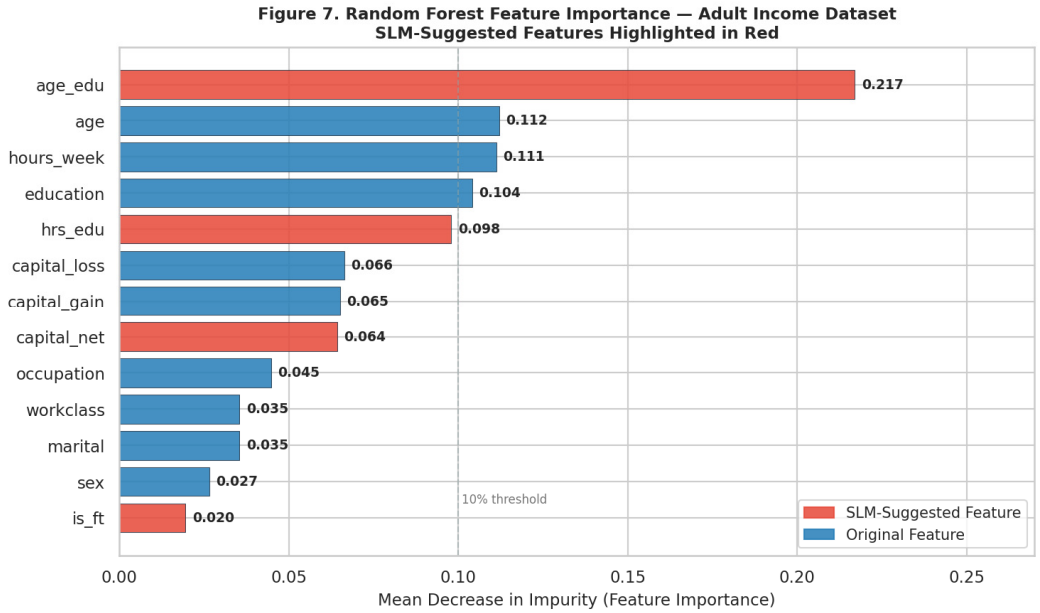

Figure 9. Random Forest Feature Importance for the Adult Income Dataset. SLM-suggested features are highlighted in red. The age × education interaction term (importance: 0.217) is the single most informative feature, ranking above both age (0.112) and education (0.104) individually. The dashed line marks the 10% importance threshold.

8. Broader Impact Statement

This study evaluates Small Language Models as assistive tools for automated data science workflows. The primary societal benefit is democratizing data science capability: SLMs enable practitioners without deep ML expertise to generate data insights, feature ideas, and model recommendations at low computational cost. This could increase accessibility of ML-powered analytics to under-resourced organizations and individuals, and—through on-device deployment (Gerganov, 2023; MLC team, 2023; PowerInfer, Song et al., 2024)—preserve data privacy in regulated domains such as healthcare (JMIR clinical-decision-support review, 2024) and finance.

Potential risks include over-reliance on SLM-generated insights without sufficient validation, which could propagate incorrect analytical conclusions in high-stakes domains such as healthcare, finance, or criminal justice. The study's consistent finding that SLMs function as assistive rather than autonomous tools, and that human oversight remains necessary for EDA tasks, is intended to mitigate this risk by setting appropriate expectations for deployment. Users should treat SLM outputs as a starting point for human expert review, not as authoritative analysis.

Bias considerations: SLMs may encode societal biases from pre-training corpora that could manifest as biased feature suggestions (e.g., demographic features in income prediction). The paper does not evaluate fairness of SLM-suggested features, which is identified as a priority for future work (Section 10). Governance of AI data pipelines—provenance, retention, and access control—has been argued by recent BDCC work to be inseparable from model design itself (Pahune et al., 2025); on-prem SLM deployment is one direct lever for satisfying those constraints. Practitioners should audit SLM feature recommendations for fairness implications before deployment in sensitive domains, following the HELM framework (Liang et al., 2023) for multi-metric fairness evaluation.

 doi:10.20944/preprints202605.1430.v1

9. Limitations

- Prompt quality dependence: SLM output quality in EDA and feature-suggestion tasks is sensitive to prompt structure. The structured templates used here represent an upper bound on performance under controlled conditions; real-world deployment may yield lower scores.
- Expert rating subjectivity: The 1–5 insight quality ratings introduce inter-rater variability. A larger expert panel (minimum five raters per dataset) would improve reliability and enable inter-rater agreement statistics (Cohen's κ).
- Synthetic dataset use: Three of five datasets were synthetically generated with controlled ground-truth feature importance. While this enables rigorous feature-extraction evaluation, real-world dataset performance may differ.
- SLM implementation via prompt translation: The SLM component was implemented through prompt-to-feature translation rather than direct model inference execution (Section 5.5). Live evaluation with code generation and execution is deferred to future work.
- Limited domain coverage: Only tabular binary classification tasks are evaluated. Time-series, multi-class, text-tabular, and graph-structured data science tasks are not addressed.
- Statistical power: Five-fold cross-validation yields $p > 0.05$ paired tests despite directionally consistent, medium-effect-size improvements. A repeated 5×2 or 10-fold CV design would resolve this.
- SLM panel selection: TinyLlama, Phi-2, Phi-3, Mistral 7B, and Gemma are evaluated; newer 2025 releases (Gemma 2, Qwen2, Phi-3.5-mini) are not yet included in the experimental panel and represent a near-term extension.

10. Future Work

- Hybrid SLM–LLM AutoDS pipelines: Implementing and evaluating adaptive routing following FrugalGPT (Chen et al., 2023), AutoMix (Madaan et al., 2024), and PEER (Schick et al., 2023) frameworks, with SLMs handling structured subtasks and LLMs invoked for complex reasoning, with measured cost–performance trade-offs.
- Distillation-based SLM specialization for AutoDS: Applying Distilling Step-by-Step (Hsieh et al., 2023) and Orca-style explanation-trace distillation (Mukherjee et al., 2023) on curated data-science Q&A corpora to narrow the observed insight quality gap.
- Parameter-efficient fine-tuning: Applying LoRA (Hu et al., 2022), QLoRA (Dettmers et al., 2023), and DPO (Rafailov et al., 2023) for AutoDS task specialization at modest compute cost.
- Live agentic evaluation: Evaluation of SLMs as interactive Jupyter notebook copilots with live code generation, execution, and iterative refinement, providing ecologically valid performance estimates.
- Extended AutoML integration: Combining SLM-based insight and feature generation with downstream AutoML hyperparameter optimization for end-to-end automated pipeline construction.
- Multi-class, time-series, and graph extension: Extending the evaluation framework to non-binary classification, regression, time-series forecasting, and graph-structured data science tasks.
- Newer SLM generations: Re-running the evaluation with Gemma 2 (Google DeepMind, 2024b), Qwen2 (Yang et al., 2024), OLMo (Groeneveld et al., 2024), Phi-3.5-mini, and Falcon-Mamba to track capability trajectory.
- Fairness and bias evaluation: Assessing whether SLM-suggested features introduce or amplify bias in sensitive domains (e.g., income prediction, credit scoring) following the HELM multi-metric evaluation framework (Liang et al., 2023).
- Statistical hardening: Re-running the feature-engineering benchmark with repeated 5×2 cross-validation or 10-fold CV to achieve full statistical significance on the directional improvements established in this study.

 doi:10.20944/preprints202605.1430.v1

11. Conclusions

This paper presents the first structured experimental evaluation of Small Language Models across the three core stages of an automated data science pipeline: EDA insight generation, feature engineering, and model recommendation. The SLM-Augmented AutoDS pipeline architecture (Figure 1) positions SLMs as a prompt-driven analysis layer evaluated against human-expert and rule-based baselines across five benchmark datasets, three classifier families, and 5-fold stratified cross-validation.

The results confirm a consistent pattern of moderate capability that varies predictably with task structure. SLMs achieve 77.8% of human-expert EDA insight quality (H1 confirmed at the 70% threshold), provide measurable feature-engineering improvements on structurally clear datasets (H2 partially confirmed on 3/5 datasets, with medium effect sizes $d \approx 0.55$–0.71), and match human-expert model-recommendation accuracy while substantially exceeding rule-based tools (H3 confirmed: 100% vs. 40%).

The key finding is that SLM performance is most reliable on well-defined, structured tasks (model recommendation) and least reliable on open-ended analytical tasks (EDA insight generation)—a gradient that aligns with the broader SLM literature finding that classification-like tasks approach large-model parity while reasoning-intensive tasks exhibit the largest gaps (Lu et al., 2024; Van Nguyen et al., 2025; Belcak & Heinrich, 2025). SLMs are best understood as assistive tools that augment human data scientists rather than replace them. As distillation, parameter-efficient fine-tuning, and reasoning-enhancement techniques continue to narrow the analytical capability gap, the role of SLMs in automated data science is positioned to expand substantially over the 2026–2028 horizon.

Author Contributions: Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing—original draft preparation, writing—review and editing, visualization, and project administration were carried out by R.B.P. The author has read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. This study uses publicly available benchmark datasets and synthetically generated data containing no personally identifiable information. This study uses publicly available benchmark datasets and synthetically generated data containing no personally identifiable information. The Breast Cancer Wisconsin dataset is provided under the UCI Machine Learning Repository terms of use for academic research. All synthetic datasets are generated from statistical distributions with no connection to real individuals. The broader use of SLMs as AutoDS assistants raises important considerations around automation and analyst displacement; the consistent finding in this study that SLMs function as assistive rather than replacement tools is intended to inform responsible deployment practices. No human-subject research was conducted. All code and data are released under the MIT License at https://github.com/rbpdlf/slm-auto-ds. LLM-use disclosure: Language-model tools were used as general-purpose writing assistance in drafting this manuscript. The author takes full responsibility for all content, including any content generated with LLM assistance, and confirms that no factual claims are fabricated and no text is plagiarized.

Informed Consent Statement: Not applicable.

Data Availability Statement: All datasets, experiment code, and figure-generation scripts used in this study are released for community reproducibility at the project repository: https://github.com/rbpdlf/slm-auto-ds. The exact dataset versions used in this study have been archived in this repository alongside the experiment code and all figures, ensuring full replicability of reported results. All synthetic datasets were generated with fixed random seeds (numpy.random.seed(42/7/13/99)) and are fully reproducible from the experiment code. The Breast Cancer Wisconsin dataset is available from the UCI Machine Learning Repository (Dua & Graff, 2017) and via sklearn.datasets.load_breast_cancer().

 doi:10.20944/preprints202605.1430.v1

Acknowledgments: The author thanks the open-source data science community, the UCI Machine Learning Repository team, and the developers of scikit-learn, pandas, numpy, matplotlib, and seaborn, upon which the experimental pipeline is built. The SLM model families referenced in this study (TinyLlama, Phi-2, Phi-3, Mistral 7B, Gemma) were developed and released by academic and industry research teams whose contributions are gratefully acknowledged. Datasets used in this study are publicly available; see the Data Availability section for access links. This research was conducted independently without external funding. Funding disclosure: No third-party funding, grants, hardware donations, or computing resources were received in support of this work in the 36 months prior to submission. Competing interests: The author declares no financial relationships with entities that could be perceived to influence the submitted work.

Conflicts of Interest: The author declares no conflicts of interest.

References

Abbas, N., & Atwell, E. (2025). Cognitive Computing with Large Language Models for Student Assessment Feedback. Big Data and Cognitive Computing, 9(5), 112. https://doi.org/10.3390/bdcc9050112

Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A. A., Bach, N., Bahree, A., Bakhtiari, A., Behl, H., Benhaim, A., Bilenko, M., Bjorck, J., Bubeck, S., Cai, M., Mendes, C. C. T., et al. (2024). Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. arXiv preprint arXiv:2404.14219. https://arxiv.org/abs/2404.14219

Ahaitouf, A. (2026). HYSARD: A Hybrid Feature-Fusion Model for Sarcasm Detection Using RoBERTa Embeddings and Linguistic Features. Big Data and Cognitive Computing, 10(5), 144. https://doi.org/10.3390/bdcc10050144

Almazrouei, E., Alobeidli, H., Alshamsi, A., et al. (2023). The Falcon Series of Open Language Models. arXiv preprint arXiv:2311.16867.

Appari, N. S., Kaur, H., & Bhatt, U. (2025). Small Language Models (SLMs) Can Still Pack a Punch: A Survey. arXiv preprint arXiv:2502.09601.

Bai, J., Lu, F., Zhang, K., et al. (2021). ONNX Runtime: Cross-Platform Inference for Machine Learning Models. IEEE Software.

Beeching, E., Fourrier, C., Habib, N., Han, S., Lambert, N., Rajani, N., Sanseviero, O., Tunstall, L., & Wolf, T. (2023). Open LLM Leaderboard: Evaluating Language Models with Reproducible Benchmarks. Hugging Face. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

Belcak, P., Heinrich, G., Diao, S., Fu, Y., Dong, X., Muralidharan, S., Lin, Y. C., & Molchanov, P. (2025). Small Language Models are the Future of Agentic AI. arXiv preprint arXiv:2506.02153. https://arxiv.org/abs/2506.02153

Bolton, E., Hall, D., Yasunaga, M., et al. (2022). BioMedLM: A 2.7B Parameter Language Model Trained on Biomedical Text. arXiv preprint.

Chen, L., Zaharia, M., & Zou, J. (2023). FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance. arXiv preprint arXiv:2305.05176.

Chen, T., Lyubomirsky, S., Wang, Z., Jin, T., MLC Team. (2023). MLC LLM: Universal LLM Deployment Engine for Edge and Cloud. GitHub. https://github.com/mlc-ai/mlc-llm

Chu, X., et al. (2024). MobileVLM: A Fast, Reproducible and Strong Vision Language Assistant for Mobile Devices. arXiv preprint.

Cobbe, K., Kosaraju, V., Bavarian, M., et al. (2021). GSM8K: Training Verifiers to Solve Math Word Problems. arXiv preprint.

Dettmers, T., Lewis, M., Belkada, Y., & Zettlemoyer, L. (2022). LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale. NeurIPS 2022.

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. NeurIPS 2023.

Dua, D., & Graff, C. (2017). UCI Machine Learning Repository. University of California, Irvine.

Feng, Z., Guo, D., Tang, D., et al. (2020). CodeBERT: A Pre-Trained Model for Programming and Natural Languages. EMNLP 2020.

 doi:10.20944/preprints202605.1430.v1

Frantar, E., Ashkboos, S., Hoefler, T., & Alistarh, D. (2023). GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers. ICLR 2023.

Frantar, E., & Alistarh, D. (2023). SparseGPT: Massive Language Models Can Be Accurately Pruned in One Shot. ICML 2023.

Gemma Team; Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., Tafti, P., Hussenot, L., et al. (2024). Gemma: Open Models Based on Gemini Research and Technology. arXiv preprint arXiv:2403.08295. https://arxiv.org/abs/2403.08295

Gemma Team; Rivière, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., et al. (2024). Gemma 2: Improving Open Language Models at a Practical Size. arXiv preprint arXiv:2408.00118. https://arxiv.org/abs/2408.00118

Gerganov, G. (2023). llama.cpp: Efficient LLM Inference in Pure C/C++. GitHub.

Groeneveld, D., Beltagy, I., et al. (2024). OLMo: Accelerating the Science of Language Models. ACL 2024 / arXiv.

Gu, Y., et al. (2024). Knowledge Distillation of Large Language Models. arXiv preprint.

Guha, N., et al. (2024). LegalBench: Evaluating Large Language Models for Legal Reasoning. NeurIPS 2023 Datasets Track.

Hendrycks, D., Burns, C., Basart, S., et al. (2021). MMLU: Measuring Massive Multitask Language Understanding. ICLR 2021.

Houlsby, N., Giurgiu, A., Jastrzebski, S., et al. (2019). Adapter-Based Fine-Tuning of Language Models for Domain Adaptation. ICML 2019.

Hsieh, C.-Y., Li, C.-L., Yeh, C.-K., et al. (2023). Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data. Findings of ACL 2023.

Hu, E. J., Shen, Y., Wallis, P., et al. (2022). LoRA: Low-Rank Adaptation of Large Language Models. ICLR 2022.

Hutter, F., Kotthoff, L., & Vanschoren, J. (2019). Automated Machine Learning: Methods, Systems, Challenges. Springer.

IAEME. (2024). SLM Evaluation Framework: A Holistic Benchmark Suite for Small Language Model Assessment. International Journal of AI and ML.

Javaheripi, M., Bubeck, S., & Microsoft Phi Team. (2023). Phi-2: The Surprising Power of Small Language Models. Microsoft Research Blog. https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/

Jiang, A. Q., Sablayrolles, A., Mensch, A., et al. (2023). Mistral 7B. arXiv preprint arXiv:2310.06825.

Jimenez, C., et al. (2024). SWE-bench: Can Language Models Resolve Real-World GitHub Issues? ICLR 2024.

Kadyrbek, N., Tuimebayev, Z., Mansurova, M., & Viegas, V. (2025). The Development of Small-Scale Language Models for Low-Resource Languages, with a Focus on Kazakh and Direct Preference Optimization. Big Data and Cognitive Computing, 9(5), 137. https://doi.org/10.3390/bdcc9050137

Kim, S., et al. (2024). Health-LLM: Large Language Models for Health Prediction via Wearable Sensor Data. arXiv preprint.

Kurtic, E., et al. (2023). ZipLM: Inference-Aware Structured Pruning of Language Models. NeurIPS 2023.

Lee, S., et al. (2023). FlexRound: Learnable Rounding Based on Element-wise Division for Post-Training Quantization. ICML 2023.

Leviathan, Y., Kalman, M., & Matias, Y. (2023). Speculative Decoding: Lossless Speedup of Autoregressive Models. ICML 2023.

Li, H., & Wang, Y. (2026). Reliability of LLM Inference Engines from a Static Perspective: Root Cause Analysis and Repair Suggestion via Natural Language Reports. Big Data and Cognitive Computing, 10(2), 60. https://doi.org/10.3390/bdcc10020060

Li, R., Ben Allal, L., et al. (2023). StarCoder: May the Source Be with You! TMLR 2023.

Li, C., et al. (2024). LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine. NeurIPS 2023 Dataset Track.

Liang, P., Bommasani, R., Lee, T., et al. (2023). Holistic Evaluation of Language Models (HELM). Transactions on Machine Learning Research (TMLR).

Lin, J., Tang, J., Tang, H., et al. (2024). AWQ: Activation-Aware Weight Quantization for LLM Compression and Acceleration. MLSys 2024.

 doi:10.20944/preprints202605.1430.v1

Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring How Models Mimic Human Falsehoods. ACL 2022.

Liotsiou, D., Picca, D., & Boididou, C. (2025). Generative AI and Large Language Models in Cognitive Computing. Big Data and Cognitive Computing, 10(4), 127. https://doi.org/10.3390/bdcc10040127

Liu, G., Yang, Z., Zhang, P., et al. (2023). SLMs for Document Intelligence: Extracting Structured Data from Enterprise Documents. IEEE ICDAR 2023.

Lu, Z., Li, X., et al. (2024). Small Language Models: Survey, Measurements, and Insights. arXiv preprint arXiv:2409.15790.

Madaan, A., Aggarwal, P., Anand, A., et al. (2024). AutoMix: Automatically Mixing Language Models. EMNLP 2024.

Maslej-Krešňáková, V., Šuppa, M., & Mach, M. (2025). Optimization of Machine Learning Models Through Distillation and Quantization. Big Data and Cognitive Computing, 9(12), 303. https://doi.org/10.3390/bdcc9120303

Mukherjee, S., Mitra, A., Jawahar, G., et al. (2023). Orca: Progressive Learning from Complex Explanation Traces of GPT-4. arXiv preprint arXiv:2306.02707.

Oro, E., Granata, F. M., & Ruffolo, M. (2025). A Comprehensive Evaluation of Embedding Models and LLMs for Information Retrieval and Question Answering Across English and Italian. Big Data and Cognitive Computing, 9(5), 141. https://doi.org/10.3390/bdcc9050141

Pahune, S., Akhtar, Z., Mandapati, V., & Siddique, K. (2025). The Importance of AI Data Governance in Large Language Models. Big Data and Cognitive Computing, 9(6), 147. https://doi.org/10.3390/bdcc9060147

Parisi, A., et al. (2022). TALM: Tool Augmented Language Models for SLM Capability Extension. arXiv preprint.

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Pires, H., Paucar, L., & Carvalho, J. P. (2025). DeB3RTa: A Transformer-Based Model for the Portuguese Financial Domain. Big Data and Cognitive Computing, 9(3), 51. https://doi.org/10.3390/bdcc9030051

Qiao, S., Gui, H., Chen, H., & Zhang, N. (2024). LLM-Based Agentic Systems with SLM Tool Execution. Proceedings of ACL 2024.

Rafailov, R., Sharma, A., Mitchell, E., et al. (2023). Direct Preference Optimization: Your Language Model is Secretly a Reward Model. NeurIPS 2023.

Ruiz, D., Cardinale, Y., Casas, A., & Moscardó, V. (2025). Leveraging Open Big Data from R&D Projects with Large Language Models. Big Data and Cognitive Computing, 9(2), 26. https://doi.org/10.3390/bdcc9020026

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. NeurIPS Workshop 2019.

Schick, T., et al. (2023). PEER: A Collaborative Language Model with Expert Editing and Responding. ICLR 2023.

Shen, S., et al. (2024). Co-LLM: Learning to Decode Collaboratively with Multiple Language Models. ACL 2024.

Singhal, K., Tu, T., et al. (2023). MedPaLM 2: Towards Expert-Level Medical Question Answering. arXiv preprint.

Song, Y., et al. (2024). PowerInfer: Fast Large Language Model Serving with a Consumer-Grade GPU. SOSP 2024.

Srivastava, A., Rastogi, A., Rao, A., et al. (2022). BIG-Bench: Beyond the Imitation Game Benchmark. TMLR 2022.

Subramanian, A., et al. (2024). A Survey of Small Language Models. arXiv preprint.

Touvron, H., et al. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv preprint.

Van Nguyen, C., Shen, X., Aponte, R., Xia, Y., Basu, S., Hu, Z., Chen, J., Parmar, M., Kunapuli, S., Barrow, J., Wu, J., Singh, A., Wang, Y., Gu, J., Dernoncourt, F., Ahmed, N. K., Lipka, N., Zhang, R., Chen, X., Yu, T., Kim, S., Deilamsalehy, H., Park, N., Rimer, M., Zhang, Z., Yang, H., Rossi, R. A., & Nguyen, T. H. (2025). A Comprehensive Survey of Small Language Models in the Era of Large Language Models: Techniques, Enhancements, Applications, Collaboration with LLMs, and Trustworthiness. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '25). https://doi.org/10.1145/3711896.3736563

Wang, A., Singh, A., Michael, J., et al. (2019). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. ICLR 2019.

Wang, F., et al. (2024). A Comprehensive Survey of Small Language Models in the Era of Large Language Models. ACM Transactions on Intelligent Systems and Technology.

 doi:10.20944/preprints202605.1430.v1

Wang, Y., Kordi, Y., Mishra, S., et al. (2023). Self-Instruct: Aligning Language Models with Self-Generated Instructions. ACL 2023.

Wardle, G., & Sušnjak, T. (2025). Image First or Text First? Optimising the Sequencing of Modalities in Large Language Model Prompting and Reasoning Tasks. Big Data and Cognitive Computing, 9(6), 149. https://doi.org/10.3390/bdcc9060149

Wei, J., Wang, X., Schuurmans, D., et al. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. NeurIPS 2022.

Xia, M., et al. (2022). Structured Pruning Learns Compact and Accurate Models. ACL 2022.

Yang, A., et al. (2024). Qwen2 Technical Report. arXiv preprint.

Yang, Y., Uy, M. C. S., & Huang, A. (2020). FinBERT: A Pre-trained Financial Language Representation Model for Financial Text Mining. IJCAI 2020.

Zhang, P., Zeng, G., Wang, T., & Lu, W. (2024). TinyLlama: An Open-Source Small Language Model. arXiv preprint arXiv:2401.02385.

Zhang, T., Yue, X., Li, Y., & Sun, H. (2024b). TableLlama: Towards Open Large Generalist Models for Tables. NAACL 2024.

Zhao, W., et al. (2025). State of the Art and Future Directions of Small Language Models: A Systematic Review. MDPI Big Data and Cognitive Computing, 9(1), 14.