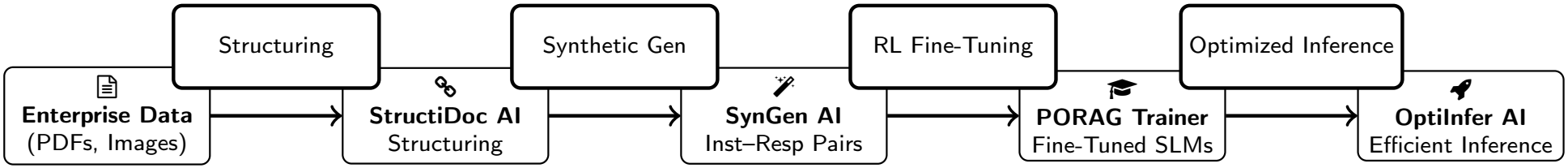
AI-Powered Enterprise Innovations

S. S. Srinivas, Shivam Gupta, Akash Das, Venkataramana Runkana

May 28, 2025

- **Foundational Capability: End-to-End AI Pipeline for Document Intelligence and Model Optimization**
  - **StructiDoc + SynGen + PORAG + OptiInfer:** A modular AI workflow that transforms unstructured enterprise documents into structured knowledge, generates synthetic instruction–response datasets, fine-tunes small language models via policy-optimized retrieval-augmented training, and deploys them with high-speed inference-time optimization.
- **Domain-Specific Application I: Automation in Chemical Engineering**
  - **AutoChemSchematic AI:** A closed-loop, physics-aware agentic framework for the automated generation and validation of chemical Process Flow Diagrams (PFDs) and Piping & Instrumentation Diagrams (PIDs) for novel chemical industrial production process.
- **Domain-Specific Application II: Advanced Solutions for Competitive Advertising**
  - **Agentic Multimodal AI for Advertising:** A framework for hyper-personalized B2B/B2C competitive advertising, leveraging multimodal market intelligence, persona simulation, and adaptive ad generation.
- **Domain-Specific Application III: Transforming B2B Chemical Commerce**
  - **ChemConnect AI:** Digitizing chemical commerce via an agentic multimodal B2B marketplace, featuring AI-driven product data structuring and actionable competitive insights.

Unified AI Workflow: From Raw Enterprise Data to Optimized AI Outputs



Stages:

- **StructiDoc AI:** Parses unstructured content into structured formats
- **SynGen AI:** Generates synthetic instruction–response pairs
- **PORAG Trainer:** Fine-tunes small language models for RAG
- **OptiInfer AI:** Enables fast, scalable, inference-time optimization

StructiDoc AI – Document Ingestion as a Service

Unlocking the Power of Unstructured Enterprise Data into Machine-Interpretable Knowledge

S. S. Srinivas, Shivam Gupta, Akash Das, Venkataramana Runkana

May 28, 2025

- **The Challenge: Proliferation of Unstructured Data**
  - Most enterprise data (e.g., scanned PDFs, contracts, invoices, spreadsheets, handwritten notes) is unstructured or semi-structured.
  - Extracting structured, machine-readable data from unstructured or semi-structured data sources remains a major bottleneck for AI-powered automation, analytics, and decision-making.
- **The Role of Structured Data in Enhancing LLM Performance:**
  - LLMs achieve factual consistency, higher accuracy, and reliability when grounded in structured, high-quality data.
  - Production AI systems demand structured inputs—yet most enterprise data is unstructured. Bridging this gap requires automated pipelines to parse, validate, and transform raw enterprise documents at scale.
- **Growing Demand for AI-Ready Data**
  - AI-native enterprises increasingly require clean, structured, and explainable data:
    - Fine-tune small-scale (LMs) on domain-specific corpora for task-specific customization
    - Build Retrieval-Augmented Generation (RAG) systems
    - Power document search, summarization, and reasoning agents
- **Our Proposed Solution:** A document ingestion platform designed to convert complex, unstructured documents into structured, machine-readable data optimized for LLMs workflows.

⊘ Limitations of Traditional OCR/RPA vs. LLM/VLM Solutions

- ⚠ **Challenges with Complex Layouts in Unstructured Documents**:
  - Multi-column formats, nested tables, and embedded visuals(Figures, Images, etc) often lead to misinterpretation or data loss.

    Example:
    - 📄 A **research paper** with 2 columns is extracted as a jumbled text stream, mixing left and right column content.
    - ▦ A **nested table** in an invoice is read row-by-row, losing column associations (e.g., "Quantity" vs. "Unit Price").

    ✔ AI Solution: Layout-aware multimodal models preserve structure in multi-column text, tables, and visuals.
- 👁 **Limited Contextual Understanding**:
  - Traditional OCR (e.g., Tesseract OCR, PaddleOCR) extracts text but cannot infer meaning or relationships.

    Example:
    - 🍎* Extracts "Apple" but cannot tell if it's **fruit** or **brand**.
    - 📄 Reads "Total: $100" and "Due: 30/05/2024" but fails to link them as part of the **same payment information**.

    ✔ AI Solution: LLMs classify "Apple" by context and group invoice fields logically.
- A **Restricted Language and Font Support**:
  - Struggles with non-Latin scripts, handwritten text, or stylized fonts.

    Example:
    - 🈶* A **Japanese Kanji** receipt is misread as random symbols.
    - ✎ **Doctor's handwritten prescription** is rendered as gibberish.

    ✔ AI Solution: Multilingual transformers (e.g., VLMs such as GPT-4o) improve accuracy across scripts.

- 🖼 **Dependence on Image Quality**:
  - Traditional OCR fails on poor scans or noisy images; requires ideal input.
    Example:
    - 🖼 Blurry ID "DL 8HX" misread as "DL 8KX" by Tesseract.

    AI Solution:
    - 👁 VLMs infer noisy text(blurred/unclear text) correctly using contextual understanding.
- 🔒 **Security and Compliance Risks**:
  - Traditional cloud OCR exposes sensitive data to third-party services.
    Example:
    - 🏥 Hospital records processed on external servers violate HIPAA(Health Insurance Portability and Accountability Act).

    AI Solution:
    - 🛡 On-device OCR systems and language-only AI models enable secure document understanding tasks—such as summarization, Q&A, etc—without relying on external servers.

- ⚙ **Inflexibility**:
  - Robotic Process Automation(RPA) tools (e.g., UiPath, Blue Prism) are software robots (or "bots") to automate repetitive, rule-based tasks that are typically performed by humans in digital systems.
    What RPA Tools Can Do with Documents:
    - 📄 Open a scanned PDF → run OCR → extract key fields → enter in form
    - ☑ Parse structured forms → validate values → trigger follow-up workflows
  - RPA tools are effective only when document formats are rigid and predictable.
  - But RPA tools breaks on slight variations in structure or layout.
    Example:
    - 📄 A new invoice template requires re-training the entire RPA pipeline.

    ✔ AI Solutions:
    - ⇄ Claude 3 handles 100+ invoice formats out-of-the-box with layout-agnostic reasoning to generalize across structural variations.
- 🕒 **High Costs**:
  - Traditional systems require constant maintenance
    Example:
    - 👥 Full-Time Equivalents(FTEs) needed to correct insurance claim OCR errors

    ✔ AI Solutions:
    - 🎓 Fine-tuned SLMs automate extraction, reducing the need for manual FTE intervention.

- **Vision-Centric Document Processing Engine:**
  - Uses layout-aware models to segment and classify document components—such as text blocks, tables, images, and figures—across both standard and non-standard layouts, including multi-column and nested structures
  - Applies specialized extraction pipelines for each content type (e.g., figure-caption linking), preserving semantic relationships and visual hierarchy
  - Reconstructs the document into a structured, LLM-ready format that retains its original meaning and context, enabling accurate downstream applications like RAG.
- **Advanced Document Parsing:**
  - Uses multi-pass Agentic OCR with VLMs to accurately extract structured data from complex documents. Generates machine-readable formats (`JSON`, `XML`, `HTML`) optimized for LLM pipelines and retrieval systems.
  - Supports custom schema definitions to fit domain-specific data extraction needs.
- **Deployment Flexibility:**
  - Supports both cloud-hosted SaaS and secure on-premises installations, ideal for regulated industries handling sensitive documents.
  - Provides REST APIs and Python SDKs for both synchronous and asynchronous ingestion workflows.
- **Security and Compliance:**
  - Enforces zero data retention—no documents are stored post-processing.
  - Offers air-gapped and on-premises deployment for maximum data privacy.
  - Compliant with HIPAA and SOC 2 Type 2, ensuring security and privacy controls.

☑ 1. RAG Accuracy (End-to-End QA)

- Evaluates how accurately the system extracts and interprets document content for downstream RAG tasks such as question answering.

Key Metrics:

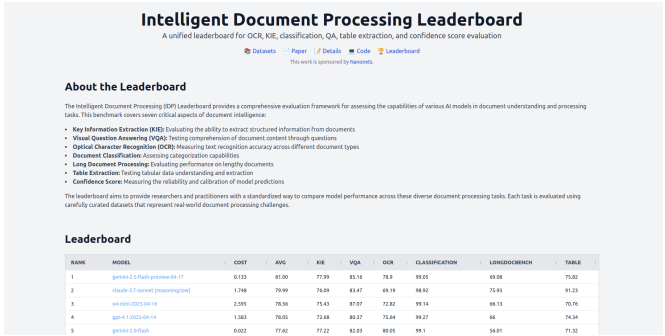
- ◎ **Exact Match (EM)** and **F1-score** on simple QA tasks.
- 📊 **BERTScore**, **ROUGE**, and **BLEU** for summarization quality.

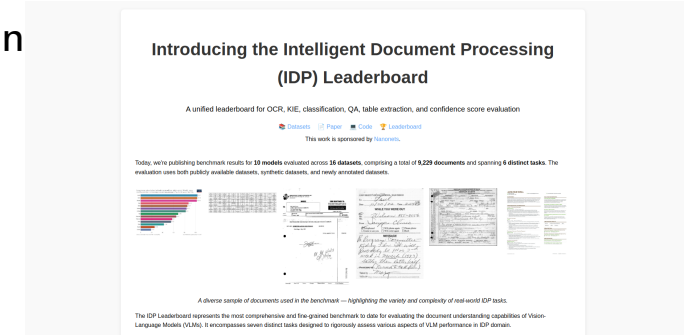
⚙ 2. Processing Efficiency

- Measures how quickly and efficiently the system processes documents at scale.

Key Metrics:

- ⏳ **Latency** (seconds per document).
- 🚀 **Throughput** (documents per second or pages per minute).
- ⚡ **Inference time** for layout and extraction modules.





Reference: Intelligent Document Processing Leaderboard
idp-leaderboard.org

SynGen AI – Synthetic Dataset Generation as a Service

Transforming Enterprise Documents into High-Fidelity Synthetic Instruction–Response Pairs for Small Language Model Training and RAG Optimization

S. S. Srinivas, Shivam Gupta, Akash Das

May 28, 2025

- StructiDoc AI transforms unstructured documents into structured data.
- Off-the-shelf SLMs are general-purpose and not fine-tuned on enterprise-specific tasks.
- SynGen AI provides high-fidelity synthetic data (QA pairs) from enterprise structured data tailored to customize SLMs for Domain-specific RAG over enterprise knowledge.

- **Synthetic Dataset Creation Workflow:**
  - **Step 1: Input Source Collection**
    - Collect enterprise documents and fed into StructiDoc AI platform, to obtain structured, machine-readable data optimized for AI workflows.
  - **Step 2: Instruction–Response Pair Generation**
    - Use large Vision-Language Models (e.g., GPT-4o, Gemini 2.5 Pro) to generate synthetic instruction–response pairs.
  - **Step 3: Dataset Typing for Specialized SLM Training**
    - **Verified Fact QA:** Extracts fact-grounded answers from complex enterprise documents.
    - **Reasoned Explanation QA:** Generates step-by-step reasoning and explanations.
    - **Contextual Grounding QA:** Produces responses grounded in specific localized or distributed content segments.
  - **Step 4: Quality Evaluation with Reward Models**
    - Filter and validate responses using reward models like Nemotron-4-340B to ensure factuality, relevance, and completeness.
  - **Step 5: Student SLM Training via PORAG**
    - Fine-tune smaller SLMs using Policy-Optimized RAG (PORAG) on these verified synthetic datasets for improved retrieval-augmented reasoning.
- **Outcome:**
  - Cost-effective, privacy-safe, and specialized training data powering high-fidelity RAG systems for enterprise document understanding.

🎓 PORAG Trainer – Policy-Optimized Fine-Tuning as a Service

🪄 Leveraging 📄 Synthetic Instruction–Response Datasets to Fine-Tune ⚙️ Specialized SLMs for Accurate RAG

S. S. Srinivas, Shivam Gupta, Akash Das

May 28, 2025

- RL fine-tuning technique to customize SLMs on synthetic instruction–response datasets generated from enterprise data, optimizing the model to generate accurate responses for domain-specific RAG.
- It is essentially to teach the SLM to effectively utilize the retrieved context and generate responses grounded in enterprise knowledge.

- **1. Inefficiency of Current Retrieval-Augmented Generation (RAG) Systems**
  - RAG pipelines power many real-world applications (e.g., search, chatbots, copilots) by grounding LLM responses with external knowledge.
  - However, they suffer from:
    - Redundant or irrelevant retrievals, increasing latency (↑ ms/s per query).
    - Unnecessary computational overhead, lowering throughput (↓ tokens/s).
    - Key-Value (KV) caching overhead, increasing GPU VRAM usage (↑ GB). This limits the maximum context length and reduces batch size, increasing memory overhead and computational cost.
  - These limitations make RAG pipelines inefficient, costly, and difficult to scale in production environments.

Why Policy-Optimized Retrieval-Augmented Generation (PORAG) Is Relevant

- **Problem:** Existing RAG systems struggle with effective utilization of retrieved context
- Group Relative Policy Optimization (GRPO) is a reinforcement learning-based fine-tuning algorithm to enhance the reasoning capabilities of LLMs.
- **Our Approach:** Fine-tune SLMs through policy optimization over retrieved contexts
  - Integrates retrieval directly into the instruction tuning process
  - The "policy" refers to the SLM's parameters that govern text generation
  - Uses GRPO to update the language-only model parameters
  - Keeps retrieval mechanism fixed (computational efficiency)
- **The GRPO Loss Function:**
  - The GRPO loss function is a clipped policy optimization objective with group-relative advantage and KL penalty.
  - **Fine-tuning Efficiency:** Uses QLoRA to reduce memory and compute overhead during training
- **Composite Reward Function for SLM Policy Optimization:**
  - Applies only to generated responses (retrieval is fixed)
  - $R(y) = 0.3 \times \text{ROUGE-L F1} + 0.2 \times \text{Length Ratio Penalty} + 0.5 \times \text{LLM-as-Judge Score}$
  - Optimizes for semantic similarity, brevity, factual correctness, and relevance
- **Key Benefits:**
  - **Efficient Inference:** Single-shot decoding with standard sampling
  - **No Multi-Candidate Ranking:** Avoids expensive reward computation at inference
  - **Superior Performance:** Significantly outperforms vanilla RAG on factuality metrics

OptiInfer AI – Test-Time Inference Optimization as a Service

Optimizing Language Model Serving for Speed, Efficiency, and Scalability

S. S. Srinivas, Shivam Gupta, Akash Das, Venkataramana Runkana

May 28, 2025

- Optimizes language model serving for faster response generation without modifying model weights.
- Reduces latency, memory usage, and cost using system-level and reasoning-level optimizations.
- Enables scalable and efficient inference for production-grade RAG applications.

- **Limitations of Static RAG:**
  - **Unnecessary Retrievals:** Always retrieves without checking if the available context is already sufficient, resulting in unnecessary latency and higher retrieval costs.
  - **Imprecise Querying:** Builds a static query from the initial user input, missing opportunities to adapt queries based on evolving context or partial answers, leading to incomplete or inaccurate responses.
  - **Fixed Reasoning Depth:** Uses static generation lengths, risking over-generation on simple tasks or under-generation on complex tasks.

- **Adaptive Inference Optimization for RAG:**
  - It modifies the behavior at inference time by dynamically deciding when to retrieve and what to retrieve based on the evolving context during generation without altering the model weights.
  - **Dynamic Retrieval:** Retrieves only when the technique detects gaps in knowledge or high uncertainty.
  - **Context-Aware Querying:** Leverages attention over the entire context to build precise, context-aware queries targeting missing information to fill information gaps to generate accurate response.
  - **Adaptive Reasoning:** Varies generation depth based on task complexity and evolving context, balancing quality and efficiency.

.

- **Key Advantage:**
  - Improves retrieval precision, reduces latency, and enhances factuality.

- We focus on low-level system-level optimizations that improve hardware-level performance to maximize runtime performance of SLMs.
- System-level optimizations focus on improving runtime efficiency of language models without modifying their parameters, targeting key system-level metrics: latency, throughput, and memory usage.

Key Performance Metrics:

- **Latency:** Time taken to generate a complete response (lower is better)
- **Throughput:** Number of tokens generated per second (higher is better)
- **Memory Efficiency:** GPU memory (VRAM) consumption impacting batch size and scalability

Techniques:

- **FlashAttention:** Efficient attention computation reduces memory bandwidth bottlenecks, improving both **latency** and **throughput**.
- **PagedAttention with KV-Cache Quantization:** Organizes the KV-cache into non-contiguous memory blocks to avoid fragmentation, improving **memory efficiency** and supporting larger batch sizes.
- **Lookahead Decoding:** Speculatively generates and verifies tokens in parallel to reduce generation latency while maintaining **output quality**.

Characteristics:

- Require **no retraining or fine-tuning** of model weights. Do not require multiple decoding passes, focus on accelerating vanilla decoding while maintaining output quality. Purely engineering/system-level improvements.

At test time, algorithmic or reasoning-level optimizations can significantly improve the **factuality**, **reliability**, and **quality** of model outputs by modifying the generation strategy—**without requiring any fine-tuning or retraining of model weights**.

Key Focus Areas:

- **Multi-Path Reasoning:** Explore multiple reasoning trajectories and select the most consistent answer to improve robustness.
- **Expert-Like Reflection:** Simulate expert behaviors such as critique, reflection, and structured re-evaluation.

Core Characteristics:

- Works entirely at inference-time without modifying model weights.
- Focuses on improving **output quality** rather than computational speed.
- May increase computational cost by generating and evaluating multiple candidate responses.
- Relies on advanced decoding algorithms rather than parameter updates or retraining.

Benefits at a Glance:

- ✔ Improve response quality without model fine-tuning.
- ✔ Enhance factuality by verifying consistency across reasoning paths.
- ✔ Dynamically control computational effort based on task complexity.
- ✔ Simulate expert-like critique and structured reasoning to improve reliability.

- **Self-Consistency:** Selects the most consistent answer by clustering multiple independently generated reasoning paths.
- **Best-of-N Sampling:** Picks the best from N candidates by self-evaluating response quality.
- **Chain-of-Thought with Reflection:** Guides reasoning through structured thinking, reflection, and answering phases in a single pass.
- **Entropy-Guided Decoding:** Dynamically adjusts sampling parameters based on model uncertainty to balance exploration and precision.
- **Chain-of-Thought Decoding:** Explores multiple reasoning paths and selects the most reliable based on token-level scoring.
- **RE^2 (Re-Reading and Re-Analyzing):** Structures reasoning into reading, re-reading, and final answer phases for deeper analysis.
- **Mixture of Agents (MoA):** Combines diverse generation, critique, and synthesis to produce refined responses.
- **Reimplementation Then Optimize (RTO):** Refines solutions by re-implementing from extracted specs and optimizing the final output.
- **PlanSearch:** Decomposes complex queries into multi-step planning and transformation stages before answering.
- **Monte Carlo Tree Search (MCTS):** Searches through reasoning paths using simulation and backpropagation for optimal responses.
- **R* Algorithm:** Uses guided tree search with consistency checks to ensure reliable and structured reasoning.

- **Policy-Optimized RAG (PORAG):**
  - RL fine-tuning technique for domain customization of SLMs
  - Significantly improves factual accuracy in responses grounded in enterprise knowledge.
- **Adaptive Inference for RAG:**
  - Decides when to retrieve based on uncertainty or knowledge gaps during generation.
  - Decides what to retrieve by building precise, context-aware queries targeting only missing information.
  - Balances retrieval effort and response quality to reduce latency and cost without retraining the model.
- **Inference-Time Optimizations:**
  - Accelerates token generation and reduces memory usage
  - Improves output quality through multi-path verification
  - Enhances reliability through expert-like reasoning patterns
  - Achieves better results without the need for fine-tuning

Key Impact: Enables faster, more accurate, and cost-efficient RAG across diverse applications and environments.

- **Unlocking Value Across the Full AI Stack**
  - **Data Layer:** Automate extraction, structuring, and synthetic data generation from enterprise documents.
  - **Model Layer:** Fine-tune SLMs with PORAG for domain-specific RAG optimization.
  - **Inference Layer:** Deploy scalable, cost-effective model serving with OptiInfer AI.

Data Preparation Pipeline

Unstructured Docs PDFs, Images $\xrightarrow{\text{StructiDoc}}$ **Structured Data** JSON, Graphs $\xrightarrow{\text{SynGen}}$ **Synthetic Data** Inst-Resp Pairs

Model Training and Inference Pipeline

Synthetic Data Inst-Resp Pairs $\xrightarrow{\text{PORAG}}$ **Fine-Tuned SLMs** RAG-Ready Models $\xrightarrow{\text{OptiInfer}}$ **AI Responses** Accurate, Fast

AutoChemSchematic AI: A Closed-Loop, Physics-Aware Agentic Framework for Auto-Generating Chemical Process and Instrumentation Diagrams

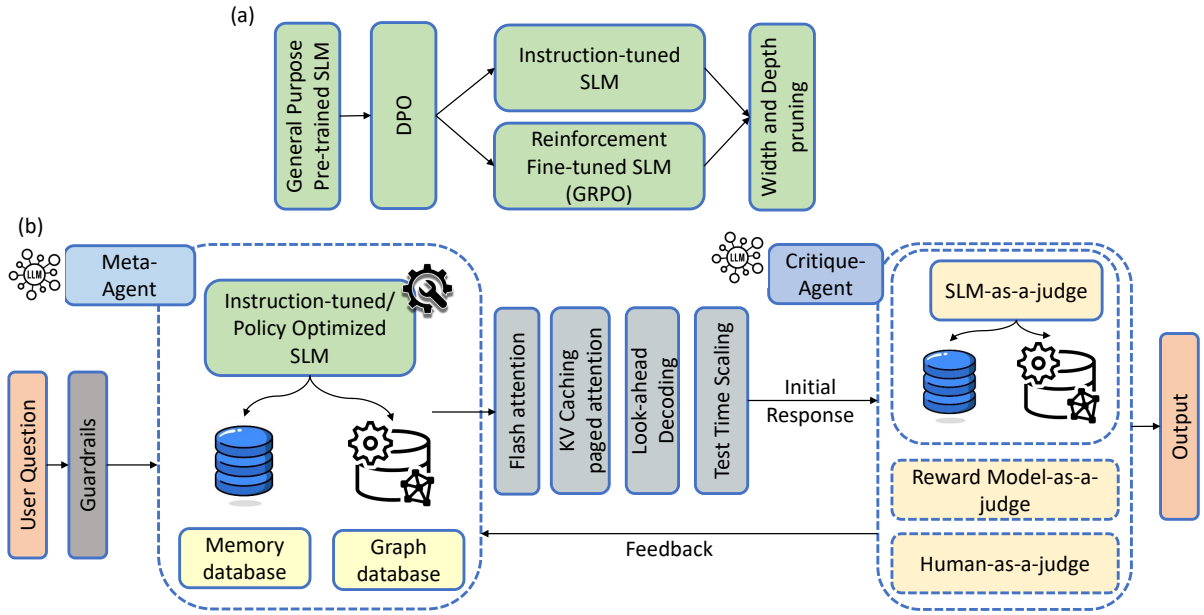
S. S. Srinivas, Shivam Gupta, Akash Das, Venkataramana Runkana

May 28, 2025

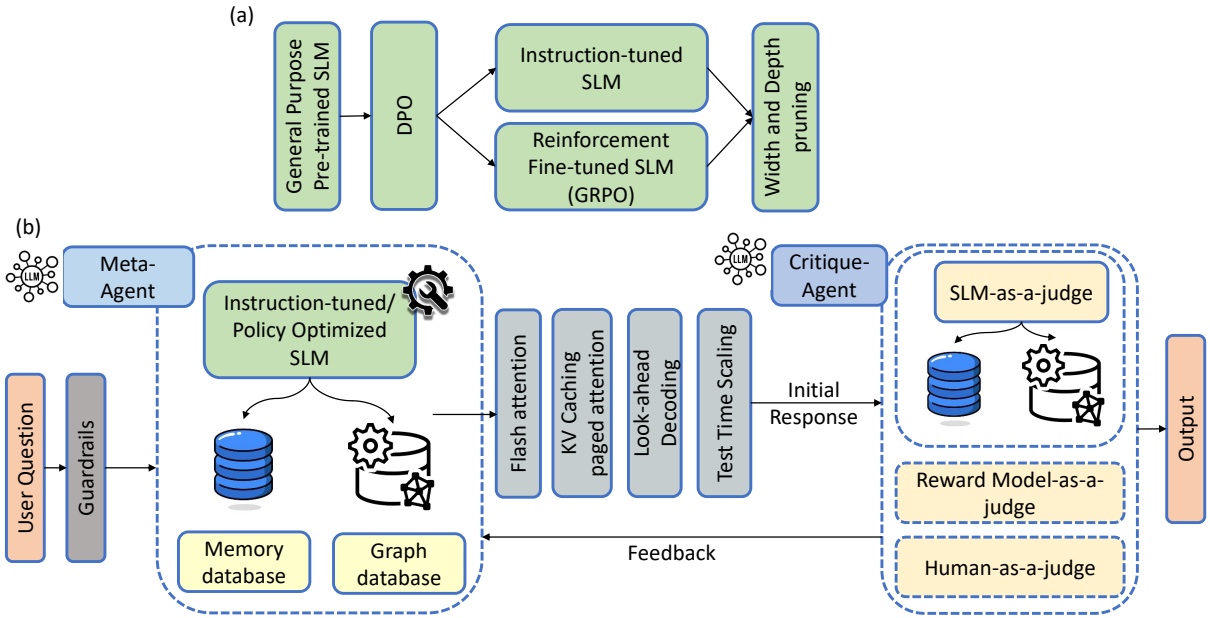
- **Process Flow Diagrams (PFDs):**
  - High-level schematic showing material and energy flows through production processes.
  - Depict major equipment, process streams, and key operating conditions.
  - Highlight *what happens* and *where it happens* in the process.
- **Piping and Instrumentation Diagrams (PIDs):**
  - Build upon PFDs by detailing valves, sensors, control loops, and actuators.
  - Illustrate *how the process operates and is controlled*.
  - Essential for safety, operational stability, and maintenance.
- **Foundation for Digital Twins and AI-Driven Automation**
  - PFDs and PIDs serve as foundational engineering schematics for digital twins.
  - Digital twins integrate:
    - First-principles or data-driven methods.
    - Real-time sensor and actuator data streams from the physical process.
  - Enable dynamic monitoring, predictive control, and closed-loop optimization.
  - Power AI-driven automation across chemical manufacturing operations.
- **AI Transforming Chemicals and Materials Discovery**
  - Generative AI accelerates discovery of:
    - Environmentally sustainable specialty chemicals.
    - Low-cost, high-performance materials-based products.
  - Reduces dependency on expensive, slow lab-based trial-and-error workflows.
  - Aids in simulation workflows and lowers R&D costs.
  - Enables faster innovation and product development cycles.

- ⚠ **Deployment Bottleneck: From Discovery to Production**
  - Challenge:Scaling discoveries from simulations or lab experiments requires developing new industrial production process.
  - Current methods fail to:
    - Auto-generate industrial-scale PFDs and PIDs.
    - Justify design and control decisions.
    - Integrate physics-aware simulations for feasibility validation.
  - Result: Slow and expertise-intensive manual workflows that limit scalability.
- ⊘ **Gaps in Process Context and Feasibility Validation**
  - Existing methods overlook:
    - High-level objectives and process sequencing.
    - Operational control, monitoring, and safety logic.
  - Lack of simulator-based feasibility checks compromises reliability.
  - These bottlenecks limits digital twin accuracy and scalable AI deployment.

- **Closed-Loop, Self-Driving Lab Framework**
  - Cloud-based SaaS platform to automate:
    - High-fidelity PFD/PID generation.
    - Design, simulation, and optimization of process schematics with minimal human input.
  - **Integrates Physics-Aware AI:**
    - Combines first-principles based process simulations with AI to ensure physical and operational feasibility.
    - Validates generated schematics through reflection with simulator-based verification .
  - Provides continuous self-improvement through AI-driven feedback loops.
  - Offers end-to-end process schematics modeling and validation.
  - Expedites the simulation-to-lab-to-pilot-to-plant scale-up pipeline.
  - Ensures that only viable, sustainable, and efficient processes advance to commercialization.



- **SLM Fine-Tuning Pipeline:**
  - Begins with Direct Preference Optimization (DPO) for alignment.
  - Followed by Instruction Tuning or Group Relative Policy Optimization (GRPO).
  - Concludes with optional width and depth pruning for efficiency.
- **Operational RAG Framework:**
  - Meta-Agent coordinates task planning and tool selection. Specialized SLM interacts with memory and graph databases for context retrieval.
  - Inference accelerated using:
    - FlashAttention: Improves throughput and latency by reducing memory bandwidth bottlenecks in attention computation.



- **Core Optimizations:**
  - Paged KV Caching: Enhances memory efficiency by reducing memory fragmentation, enabling larger batch sizes during inference.
  - Lookahead Decoding: Lowers latency by speculatively generating multiple tokens in parallel without sacrificing output quality.
  - Test-Time Scaling: Increases factual accuracy by using techniques like multi-step reasoning, self-reflection, and re-ranking during inference.
- **Iterative Response Refinement:**
  - Critique-Agent manages iterative feedback loops.
  - Uses diverse judges:
    - Nemotron-4-340B reward model, SLM-as-a-judge, Human evaluations
- **Offline-Third party Process simulations Verification.**

Agentic Multimodal AI for Hyper-Personalized B2B and B2C Advertising in Competitive Markets: An AI-Driven Competitive Advertising Framework

S. S. Srinivas, Shivam Gupta, Akash Das, Venkataramana Runkana

May 28, 2025

Industry Context:

- AI accelerates material discovery in energy, electronics, and FMCG.
- Success depends on bridging the gap between industrial scale-up and market adoption.

Commercialization Challenges:

- Weak value articulation limits market acceptance.
- New products struggle against established brands.
- One-size-fits-all messaging fails across regions.

Market Risks:

- Price wars from undifferentiated messaging.
- SKU-level cannibalization in product portfolios(Products from the same brand compete with each other, hurting overall sales).
- Low engagement from non-localized campaigns.

Limitations of Current Tools:

- Siloed R&D and marketing with no market feedback loop.
- Poor user modeling and static competitive insights.
- Lack of adaptive, privacy-safe campaign optimization.

Proposed AI-Driven Solution:

- Connect product-market fit with live market insights.
- Personalize engagement with data-driven targeting and adaptive creatives.
- Scale competitive messaging using agentic AI (MAAMS, PAG, CHPAS).

- **What is Programmatic Advertising?**
  - Automated buying and selling of digital ad space using software platforms.
  - Replaces manual media buying with real-time, algorithmic auctions across digital channels.
- **Advertising Channels in the Chemical Industry:**
  - Search: Google Ads, Bing Ads — keyword-based auctions for buyer intent.
  - Social: LinkedIn (B2B), Instagram, TikTok (B2C) — audience-based ad delivery.
  - E-commerce: Amazon, Alibaba, Knowde — product discovery and lead generation.
- **Key Stakeholders:**
  - Publisher: Platform offering ad inventory (e.g., Google, Knowde).
  - Advertiser: Chemical manufacturers promoting products to buyers.
  - SSP (Supply-Side Platform): Manages and sells publisher inventory.
  - DSP (Demand-Side Platform): Enables real-time bidding by advertisers.
- **How It Works:**
  - Powered by Real-Time Bidding (RTB)—an automated auction system for selecting the most relevant ad.
  - RTB is triggered instantly when a user visits a digital property.

- **What is RTB?**
  - RTB is the auction engine behind programmatic advertising.
  - It operates across search, social, and e-commerce channels.
  - Each ad opportunity—called an impression—is evaluated and sold in real time, typically within 100 milliseconds.
- **How RTB Works:**
  1. A user visits a digital property (e.g., Google Search, LinkedIn feed, Knowde product page).
  2. The SSP offers the impression to multiple DSPs.
  3. DSPs evaluate:
     - User signals (location, industry, behavior).
     - Page context (chemical category, product detail).
     - Channel data (search intent, social engagement).
  4. Advertisers submit real-time bids.
  5. The highest bidder wins; their ad is instantly shown.
- **Auction Types:**
  - Header Bidding: A pre-auction strategy where multiple SSPs bid simultaneously for the same impression—maximizing competition and publisher revenue.
  - Second-Price Auction: A pricing mechanism where the highest bidder wins, but pays only the second-highest bid—ensuring fairness and cost-efficiency for advertisers.

Header Bidding (Pre-Auction Strategy):

- A pre-auction mechanism where multiple SSPs bid simultaneously.
- Replaces the "waterfall" approach, where only one SSP (e.g., Google Ad Manager) gets priority.
- The highest SSP bid is passed to the final ad server auction.

Why it matters:

- Increases competition across SSPs (e.g., Google, Amazon, Xandr).
- Boosts publisher revenue and avoids walled garden dominance.

First-Price Auction:

- The winner pays exactly what they bid.
- Leads to strategic bidding, not truthful bidding—creates risk and inefficiency.

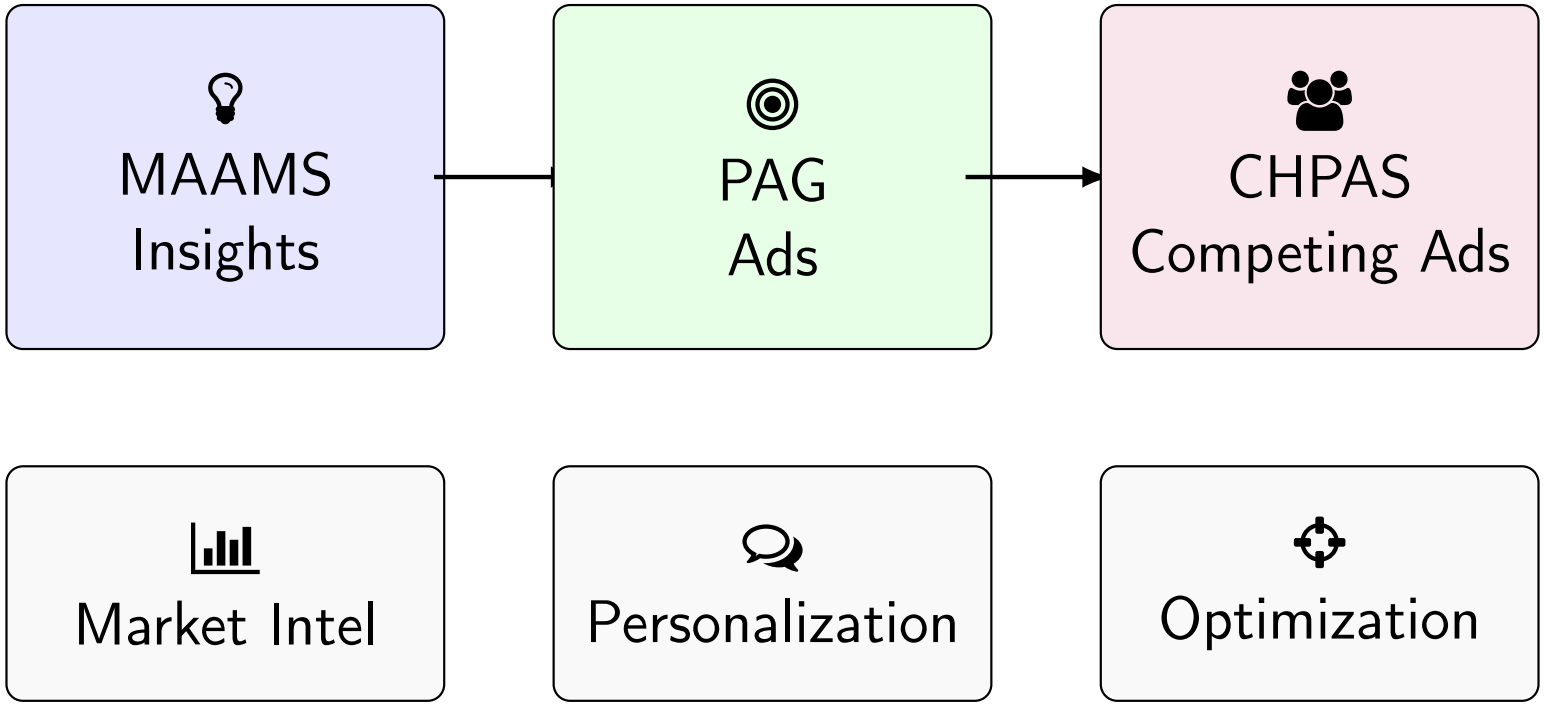
Second-Price Auction (used in RTB):

- The highest bidder wins, but pays only the second-highest bid + \$0.01.
- Encourages truthful bidding—advertisers can bid their actual maximum value.
- Reduces overpayment risk, improving auction fairness and pricing efficiency.

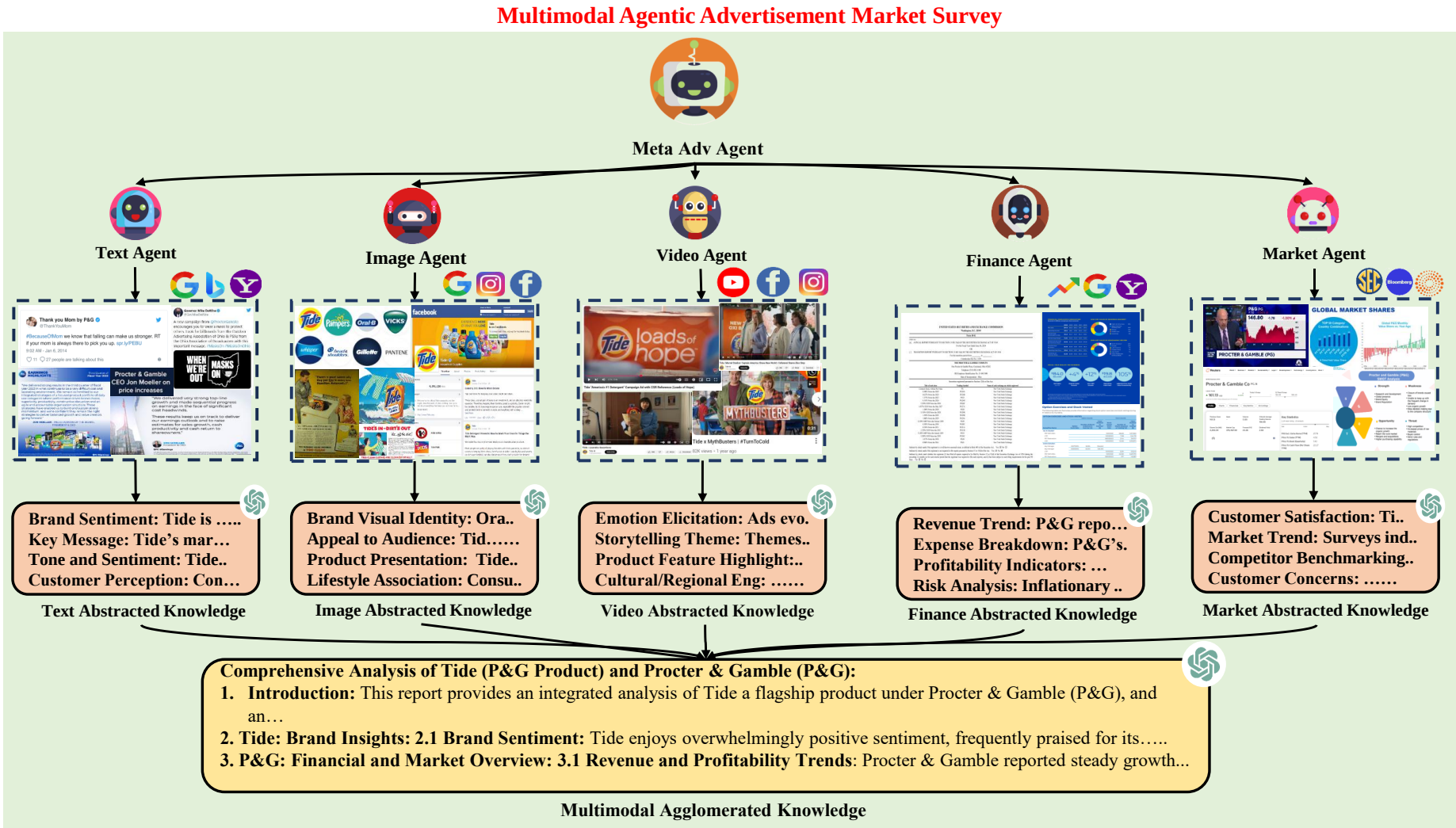
- Real-Time Bidding (RTB) selects ads in milliseconds, but effectiveness depends on targeting the right users.
- Ads shown to low-intent users waste budget—even if they win the auction.
- Precise targeting turns a fast auction into a smart investment by improving return on ad spend (ROAS).
- Data-driven targeting uses first-party, third-party, and contextual signals to identify high-intent users.
- Targeting Inputs:
  - First-Party Data: From CRM systems and website behavior.
    - Customer profiles, purchase history, product preferences.
    - Website actions like product views, downloads, sample requests.
  - Third-Party Data: From Data Management Platforms (DMPs).
    - Industry-specific segments (e.g., plant managers, chemical buyers).
    - Facility type, production capability, and procurement roles.
  - Contextual and Geographic Signals:
    - Page relevance (e.g., chemical applications, compliance).
    - Location-based filters (e.g., logistics feasibility).
    - Privacy-aware processing (GDPR, CCPA compliant).
- These inputs enable DSPs to bid intelligently—based not just on price, but on the likelihood of conversion—across search, social, and e-commerce channels.

- Winning the bid secures ad space—but performance depends on delivering a relevant and personalized creative to the user.
- ML-driven creative selection considers:
  - User attributes (profile, behavior, intent).
  - Channel context (search, social, e-commerce).
  - Historical performance (CTR, conversions).
- Format varies by channel:
  - Static banners in search.
  - Videos or carousels in social feeds.
- Dynamic Creative Optimization (DCO) personalizes:
  - Product visuals — e.g., lab equipment for scientists vs. bulk packaging for procurement teams.
  - Messaging tone — e.g., "improve yield and purity" for R&D vs. "reduce cost and downtime" for operations.
  - Highlighted benefits — e.g., regulatory compliance for pharma vs. sustainability metrics for specialty chemicals.
- Performance metrics (clicks, downloads, sample requests) feed back into future targeting and creative optimization.
- *Takeaway:* Personalizing the ad creative after winning the bid is essential—the right message, delivered to the right user, drives outcomes. Winning the bid alone is not enough.

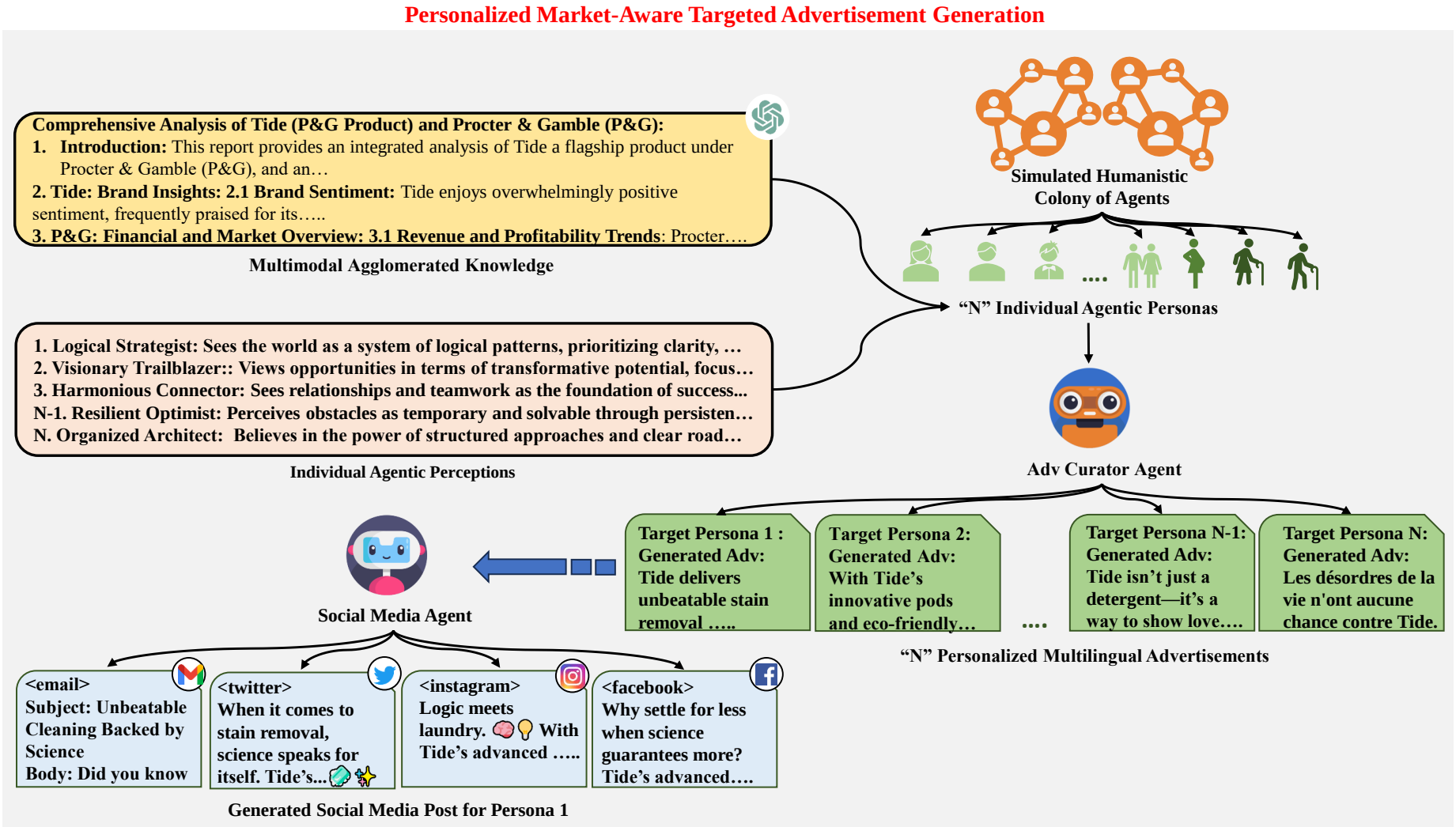
- ⚠ Limitations of Traditional Approaches:
  - Lack of dynamic creative optimization to personalize engagement at scale.
  - Weak modeling of user personas and multimodal buyer behavior.
  - Inability to differentiate similar products from competing brands or within the same portfolio.
  - Poor adaptation to privacy regulations like GDPR and CCPA in real-time targeting workflows.
- ⚙ Optimization Goals for Ad Performance:
  - Maximize Return on Ad Spend (ROAS) by optimizing targeting, messaging, and acquisition efficiency.
  - SKU (Stock Keeping Unit) cannibalization occurs when multiple products from the same company compete for the same customer, reducing total sales instead of expanding market share.
  - Ensure privacy-compliant optimization using real-time market signals.
- 📢 Why Advertising Matters in Chemical GTM(Go-To-Market):
  - Moves products beyond lab-scale innovation to real market adoption.
  - Clarifies unique technical and commercial value to non-technical buyers.
  - Helps new launches stand out from entrenched incumbents.
  - Reaches niche B2B and broad B2C audiences across global markets.
  - Drives measurable outcomes through adaptive, data-driven campaigns.



- MAAMS (Multimodal Agentic Advertisement Market Survey): Surveys the market using specialized agents to gather multimodal intelligence on brand perception, financials, and competitor positioning.
- PAG (Personalized Market-Aware Targeted Advertisement Generation): Generates personalized, multilingual ads by simulating diverse consumer personas and tailoring content to their specific preferences and cultural contexts.
- CHPAS (Competitive Hyper-Personalized Advertisement System): Optimizes these personalized ads for competitive scenarios by strategically highlighting unique selling points against rival products to maximize relevance and effectiveness for each target user.

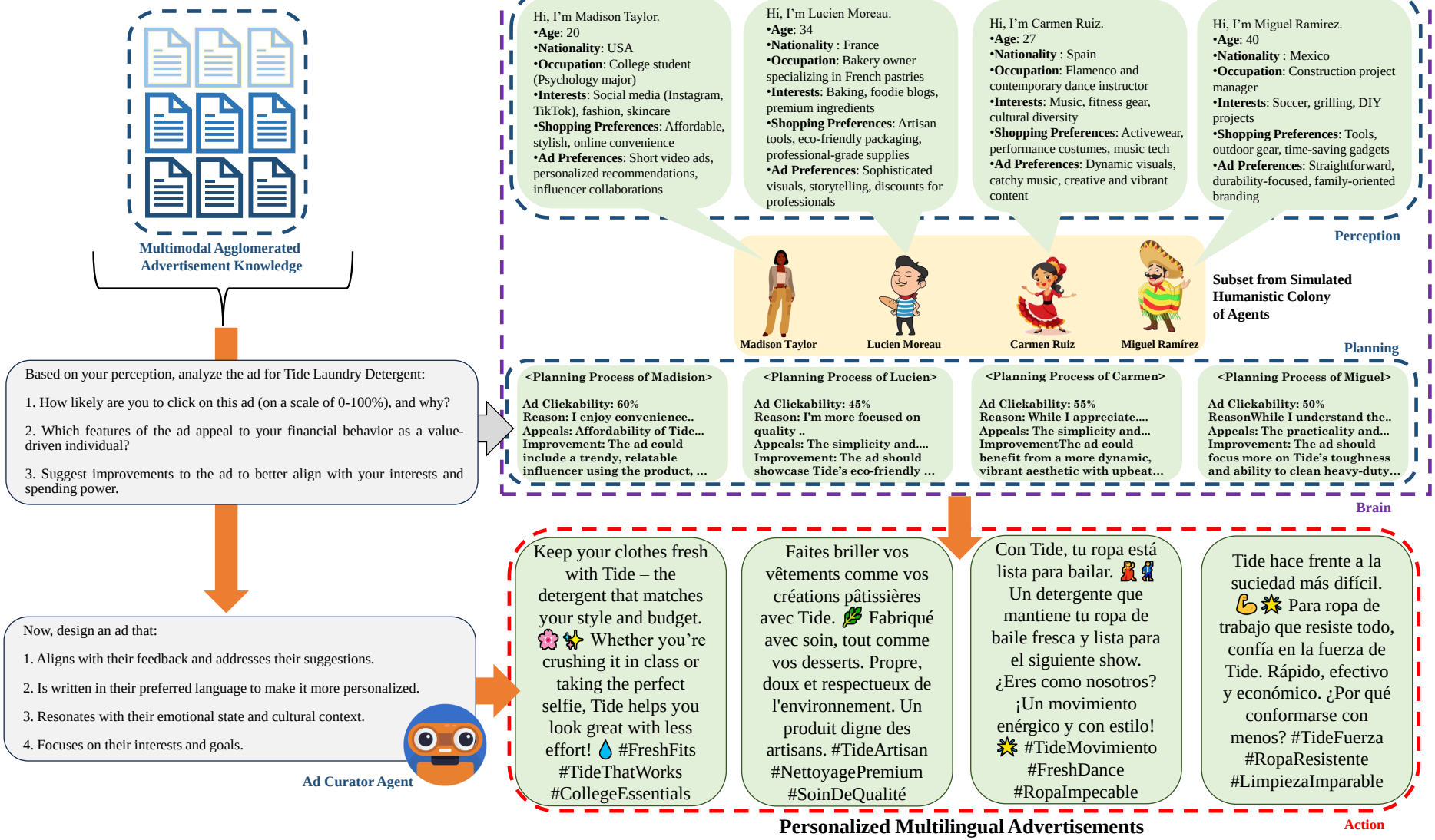


The MAAMS system aggregates multimodal insights via a Meta-Agent to analyze brand sentiment, performance, and market standing.



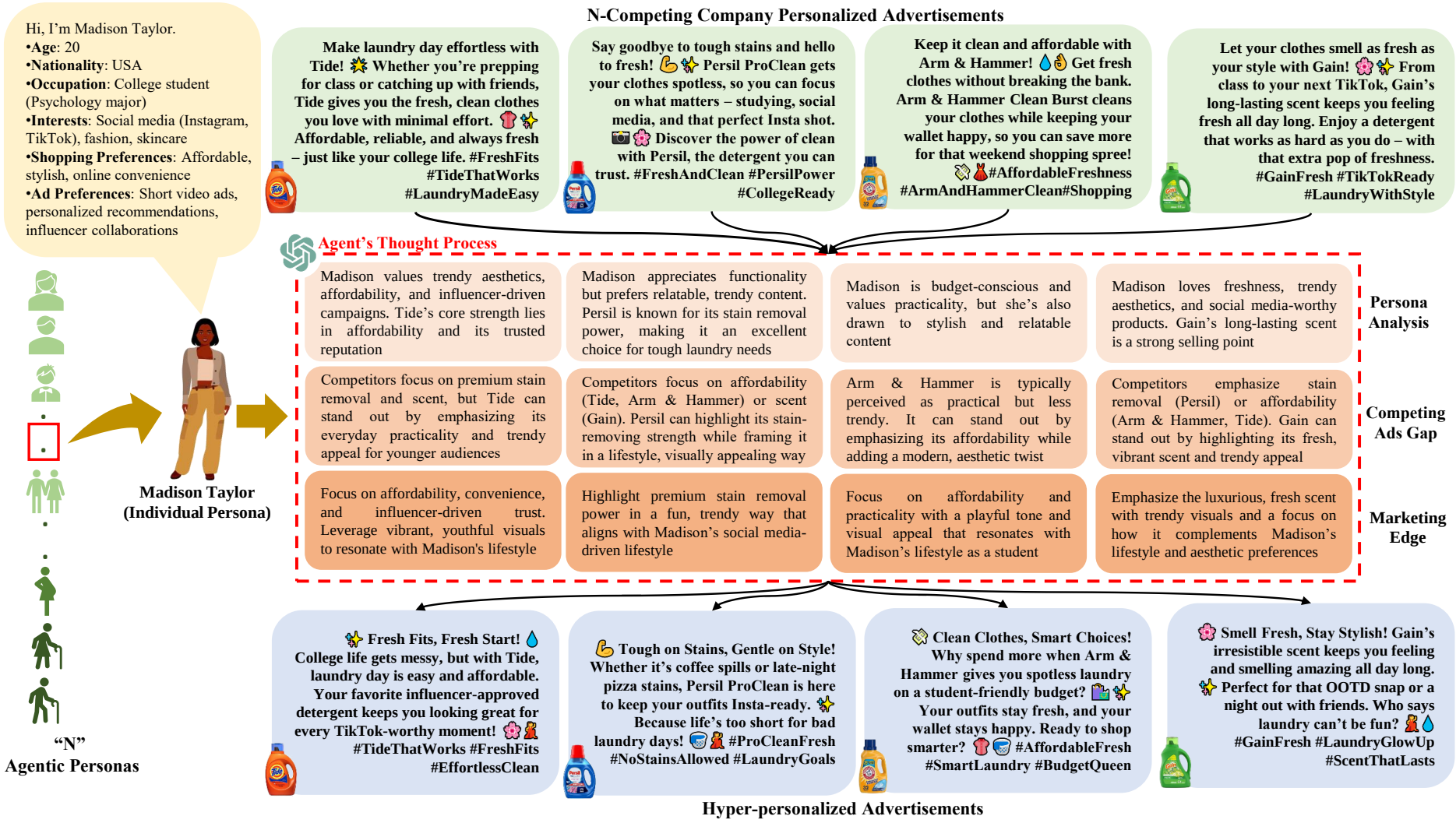
The PAG system workflow uses persona-simulating agents to generate personalized, multilingual ads, with the Adv Curator ensuring cultural fit and the Social Media Agent optimizing platform delivery.

Personalized & Multilingual Ad Creation



The PAG system generates personalized, multilingual ads by evaluating consumer preferences, cultural context, and feedback using multimodal knowledge to maximize engagement and drive action.

Hyper-Personalized Competitive Ads



Shows persona-driven, differentiated ad generation for competing products, optimizing engagement by aligning each with specific preferences and competitive strengths.

- Objective:
  - Enhance the effectiveness of AI-generated ad copy for consumer chemical products (e.g., personal care, home care, nutrition, and cleaning).
  - Maximize emotional impact, brand recall, and purchase intent across diverse consumer segments.
- Evaluation Framework:
  - Based on CHPAS: Competitive Hyper-Personalized Advertisement System.
  - Uses Simulated Humanistic Agentic Personas (SHAP) to assess ad variations across key consumer dimensions:
    - Emotional Impact Evaluator — Assesses warmth, empathy, excitement, or trust.
    - Persona Resonance Filter — Measures alignment with different lifestyle profiles (e.g., eco-conscious parent, busy professional, wellness seeker).
    - Cultural Context Checker — Ensures relevance across regions, languages, and norms.
    - Brand-Value Alignment — Validates messaging against brand positioning (e.g., sustainability, safety, luxury).
  - Scoring Dimensions (0–10 scale): Emotional Resonance, Brand Fit, Clarity, Cultural Relevance.
- Experimental Setup:
  - Run synthetic evaluations across channels like Instagram, TikTok, Meta, and e-commerce banners.
  - Test copy styles: benefit-led vs. feature-led, sensory words, visual callouts, and CTA phrasing.

Motivation: Real-world A/B testing for ad personalization is costly, risky, and constrained by privacy laws (e.g., GDPR, CCPA).

Goal: Develop a privacy-compliant framework to simulate and optimize ad strategies across competing products—entirely offline.
Pipeline Overview:

1. Persona Profiling: Generate synthetic user profiles with demographic and behavioral traits.
2. Product Modeling: Evaluate product features, pricing, and brand perception.
3. Competitive Simulation: Model competing products and market scenarios.
4. Product-Persona Alignment: Score alignment between personas and product attributes.
5. Ad Generation: Create generic and persona-optimized ads using LLMs.
6. LLM Evaluation: Score ads for relevance, persuasion, and emotional impact (e.g., GPT-4 Omni, Nemotron).
7. Ranking & Optimization: Rank ads by predicted effectiveness; iterate strategies offline.

Outcome: A scalable, cost-effective framework enabling hyper-personalized ad optimization without real-world deployment.

ChemConnect AI: Digitizing Chemical Commerce with Agentic Multimodal B2B Marketplace Intelligence

Unified Marketplace | AI Product Structuring | Data-Driven Competitive Insights

S. S. Srinivas, Shivam Gupta, Akash Das, Venkataramana Runkana

May 28, 2025

Why It Is Needed

- The $20T+ global chemical industry remains highly fragmented, with <5% of transactions conducted digitally.
- Relies on legacy mechanisms—emails, trade shows, and phone calls—that slow procurement and discovery cycles.
- Supplier data scattered across unstructured PDFs, spreadsheets, and siloed systems.
- No centralized platform for technical product discovery, structured comparison, sampling, and quoting

Why It Matters

- Improves sourcing efficiency, reducing time-to-market and costs.
- Centralizes supplier catalogs into machine-readable structured data for search and discovery.
- Digitally connects verified suppliers to high-intent global buyers.
- Accelerates R&D by simplifying material discovery, sampling, and quoting.

What Is ChemConnect AI?

- A global B2B digital marketplace for chemicals, polymers, ingredients and etc.
- Leverages AI to extract, normalize, and structure unstandardized supplier data for technical searchability.
- Provides seamless search, sampling, quoting, and procurement workflows.
  - In B2B chemical marketplaces, **sampling** refers to the process where buyers request free or paid small quantities of materials (called samples) to evaluate them before committing to bulk purchases.
  - It helps R&D teams, formulators, and procurement professionals to:
    - Test performance in their own formulations or manufacturing processes.
    - Validate quality and specifications against their requirements.
    - Compare multiple suppliers' offerings before finalizing large-scale procurement.
- Powers digital transformation across the chemical industry by connecting suppliers and buyers on a unified platform.
- Empowers suppliers to go digital through branded storefronts powered by ChemConnect's backend.

What ChemConnect AI Offers

- Marketplace: Centralized, searchable catalog of verified suppliers and products.
- AI Engine: Extracts product specs from unstructured data (PDFs, Excel).
- Master Data Management (MDM): Ensures cross-supplier consistency and normalization
- Customer Experience Platform (CXP): Powers branded storefronts with AI-driven search, filtering, and sample/quote workflows.
- ChemConnect AI offers more than just hosting products on its marketplace.
- It enables suppliers to launch their own branded digital storefronts on their company websites.
- These storefronts use ChemConnect AI's technology to provide:
  - Advanced product search and filtering.
  - Structured data presentation for easy discovery.
  - Lead capture through sample and quote requests.
  - Access to product documents, certifications, and specifications.
- This helps suppliers digitally engage customers directly on their own websites while leveraging ChemConnect AI's platform capabilities.

Stakeholders

- **Suppliers:** Upload and structure chemical products with specs, documents, and certifications.
- **Buyers:** R&D labs, formulators, and procurement teams seeking discovery, sampling, and technical procurement.

Mutual Value

- **Suppliers:** Expand reach, digitize catalogs, launch storefronts, and convert leads to deals.
- **Buyers:** Perform side-by-side comparisons, access verified data, and initiate sampling or quotes with fewer intermediaries.

Why Not Amazon, Flipkart, or Uber?

- **Consumer E-commerce:** Lacks multi-attribute, compliance-focused technical filtering.
- **Service Marketplaces:** Not designed for product procurement, sampling, or data compliance.
- **ChemConnect AI:** Purpose-built for industrial-grade search, domain-specific ontology, and technical workflows.

- **Introducing ChemConnect Insights:**
  - Subscription-based analytics platform for chemical suppliers and industrial buyers.
  - Provides real-time, sector-specific, and behavioral market intelligence to drive data-informed decisions.
- **Sector & Regional Market Intelligence:**
  - Access industry-specific demand insights across coatings, pharma, and etc.
  - Enable targeted go-to-market, pricing, and distribution strategies across global regions and sectors.
- **Real-Time Engagement Analytics:**
  - Track product views, sampling requests, quote submissions, and buyer engagement.
  - It allows suppliers to optimize product positioning, catalog performance, and customer outreach dynamically.
- **ChemConnect GPT (AI-Powered B2B Insights Assistant):**
  - AI assistant(chat-based tool) trained on ChemConnect marketplace data helps suppliers and buyers answer performance queries, generate competitive benchmarks.
  - Provides actionable sales or sourcing recommendations.
  - **Suppliers**:
    - Understand how their products are performing (e.g., view rates, quote requests, conversion rates).
    - Compare their performance to similar competitors (competitive benchmarks).
  - **Buyers**
    - Discover better or cheaper alternatives based on historical sourcing data.
    - Get recommendations on which suppliers or products fit their needs based on marketplace insights.

- **Behavioral Metrics Dashboard:**
  - Monitor buyer retention, repeat sampling behaviors, and supplier share-of-voice in in platform search and category placements.
    - Buyer Retention: How many buyers come back to the same supplier or product over time.
    - Repeat Sampling: How often the same buyers request new samples of the same product or related products.
    - Share-of-Voice: How prominently a supplier's products appear in search results and category listings compared to competitors.
  - It offers a comprehensive view of customer engagement across the marketplace.
- **$ Strategic Business Impact:**
  - Provides data-backed decision support for suppliers and buyers.
  - Unlocks a recurring SaaS revenue stream alongside marketplace transactions.
    - (i) charge a commission or fee on every successful B2B sale made through your digital marketplace.
    - (ii) Suppliers or buyers pay monthly or yearly to access advanced analytics).

We gratefully acknowledge the contributions of:

- Venkataramana Runkana
- Shivam Gupta* (B.Tech - IIT Hyderabad, and M.Tech - IIT Bombay)
- Chidaksh Ravuru (North Carolina)
- Geethan Sannidhi (Arizona)
- Sreeja Gangasani (Google)
- Vijay Sri Vaikunth (Georgia Tech)
- Akash Das* (B.Tech - IIT Patna, and M.Tech - IIT Madras)

*Current team members
The team has published 45+ works, including top-tier AI conference papers and patents, in the last 4 years alone.

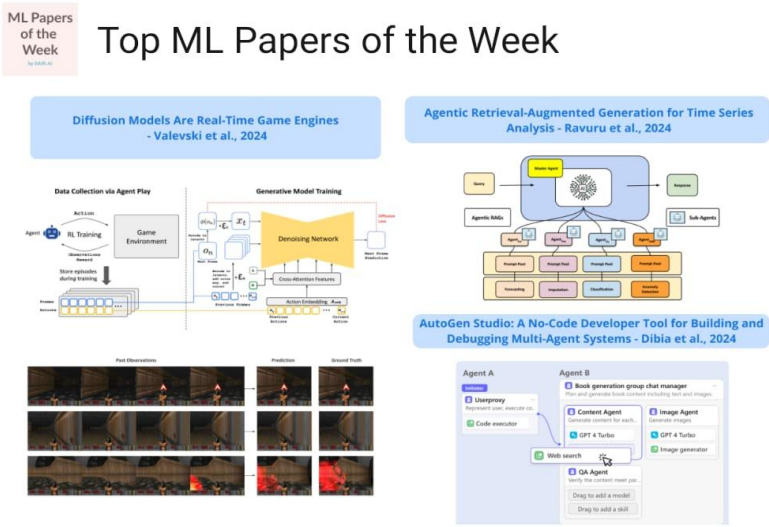
Special thanks to:

S. Iyer
Vice-President, Google Asia-Pacific

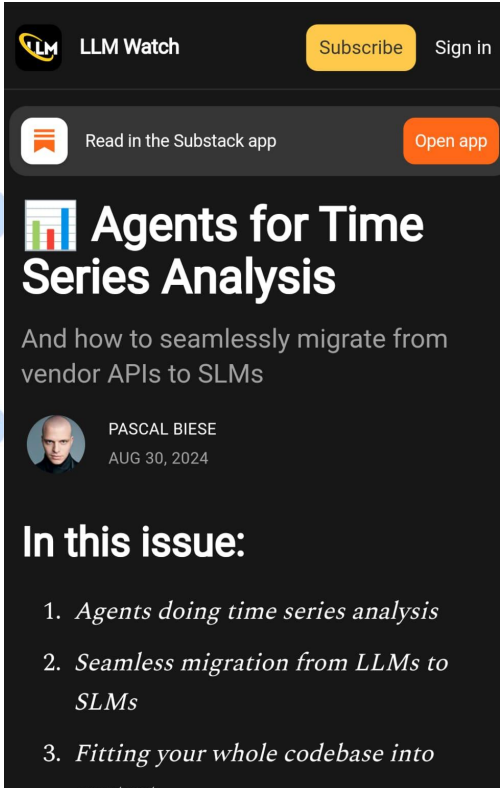
and her team members for their technical guidance, high-quality discussions, and valuable support throughout this works.

☺ *Thank you all for your collaboration and support!*

Questions & Discussion



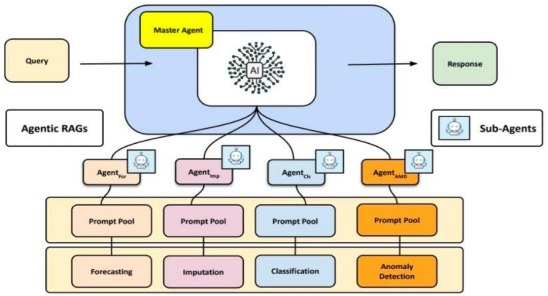
🥇 Top ML Papers of the Week



A Hierarchical Multi-Agent Framework for Enhanced Time Series Analysis

Agentic-RAG: A Hierarchical Multi-Agent Framework for Enhanced Time Series Analysis

By Sana Hassan - September 1, 2024



https://arxiv.org/abs/2408.14484

Time series modeling is vital across many fields,

Kunal Grover's Hatch

Enhancing Time Series Analysis Through Agentic-RAG and the AI Agents Stack

Hatched by **Kunal Grover**
Jan 31, 2025

4 min read · 17 views

0

Enhancing Time Series Analysis Through Agentic-RAG and the AI Agents Stack

In today's rapidly evolving technological landscape, the integration of artificial intelligence (AI) into various domains has transformed the way we approach data analysis, particularly in the realm of time series data. The emergence of frameworks like

sophisticated approach to

gents stack, these

the accuracy and reliability

George Campbell · 3rd+
Senior Account Executive | Sales and Business Development …
2w ·

+ Follow

AI affects us all differently. Yesterday, I had a moment while I finished this paper; a year ago, honestly, AI was overwhelming AF. I was deep in the weeds, reading, testing, chasing trends…the scale, the speed, the potential—it can get out of control. At least for me it did. However, everything changed when I started applying AI to what I know best: Advertising, marketing, sales…big thanks to Sean @ AI Superpowers 🙏

Anyways, I realized it's not about knowing everything. It's about seeing how AI enhances the systems you already live and breathe. So, if you're still overwhelmed, don't be embarrassed-this stuff moves nonstop, you can get lost.

Here's my (limited but real) advice: Start by applying AI to your own world—your job, your craft, your industry. Once you do that, the rest just falls into place.

It took me a few days to read 😅 but trust me: The deeper the understanding, the more fascinating it gets. Check this paper out … ⚠️ not for the overwhelmed. https://lnkd.in/eAD7f-Z5 …wicked site too-thanks Cornell.

arXiv
Agentic Multimodal AI for Hyperpersonalized B2B and B2C Advertising in Competitive Markets: An AI-…
arxiv.org

12 4 comments

Reactions

Like Comment Repost S

← Post

TuringPost ✓
@TheTuringPost

16 new research on inference-time scaling

13 new methods:

- Inference-Time Scaling for Generalist Reward Modeling
- T1: Tool-integrated Self-verification
- Z1: Efficient Test-time Scaling with Code
- GenPRM
- Can Test-Time Scaling Improve World Foundation Model?
- Scaling RAG Systems With Inference-Time Compute Via Multi-Criteria Reranking
- φ-Decoding
- Inference-Time Scaling for Flow Models
- Dedicated Feedback and Edit Models Empower Inference-Time Scaling for Open-Ended General-Domain Tasks
- m1: Test-Time Scaling for Medical Reasoning
- ToolACE-R
- Scaling Test-Time Inference with Policy-Optimized, Dynamic Retrieval-Augmented Generation via KV Caching and Decoding

ime Compute

ne Scaling in

PAPER REVIEW

Paper Review: Agentic Retrieval-Augmented Generation for Time Series Analysis

Time-Series predictions, now with LLM and RAG

Andrew Lukyanenko Follow 3 min read · Sep 4, 2024

18

Paper link

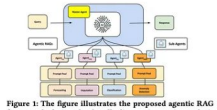
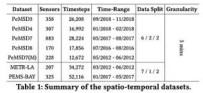


Figure 1: The figure illustrates the proposed agentic RAG framework, designed to handle diverse time series analysis tasks. The framework employs a hierarchical, multi-agent architecture. A master agent receives end-user questions and routes them to appropriate specialized sub-agents based on the specific time series task (e.g., forecasting, imputation, classification, anomaly detection). The sub-agents utilize pre-trained SLMs fine-tuned on task-specific datasets using techniques like instruction tuning and direct preference optimization to capture spatio-temporal dependencies within and across the time series datasets. Each sub-agent main-



Dataset	Sensors	Timesteps	Time-Range	Data Split	Granularity
PeMSD3	358	26,208	09/2018 - 11/2018	6 / 2 / 2	5 MIN
PeMSD4	307	16,992	01/2018 - 02/2018		
PeMSD7	883	28,224	05/2017 - 08/2017		
PeMSD8	170	17,856	07/2016 - 08/2016		
PeMSD7(M)	228	12,672	05/2012 - 06/2012	7 / 1 / 2	5 MIN
METR-LA	207	34,272	03/2012 - 06/2012		
PEMS-BAY	325	52,116	01/2017 - 05/2017		

Table 1: Summary of the spatio-temporal datasets.

Dataset	SWaT	WADI	SMAP	MSL	TEP	HAI
Sensors	51	123	25	55	52	59
r	25	25	50	55	35	30

Table 2: Statistical summary of benchmark datasets. r is the length of subsequences or historical window length.

The Deep Hub

Your data science hub. A Medium publication dedicated to exchanging ideas and empowering your knowledge.

Follow publication

Agentic RAG Explained: AI's New Approach to Time Series Problems

Malyaj Mishra Follow 4 min read · Dec 15, 2024

Share

99 2

2024-09-01

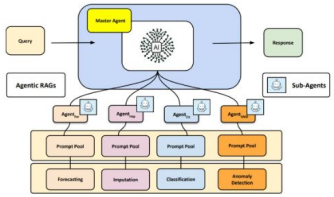
Agentic Rag Pipeline For Time Series Analysis

Agentic Retrieval-Augmented Generation for Time Series Analysis

Chidaksh Ravuru
IIT Dharwad
India
20001004@iitdh.ac.in

Sagar Srinivas Sakhinana
TCS Research
India
sagar.sakhinana@tcs.com

Venkataramana Runkana
TCS Research
India
venkat.runkana@tcs.com



Agentic RAG (Retrieval-Augmented Generation) is quickly becoming a hot topic in AI. And now, we're about to witness another level up—Agentic RAG, especially for time-series analysis.

So, what is it? And why should you care?

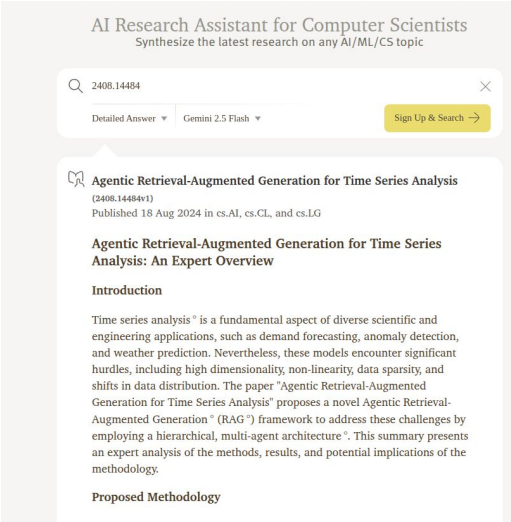
What is Agentic RAG?

Imagine you're managing a huge project, and you've got a team of specialists working under you—each person assigned to a specific

I also explore innovative approaches to Retrieval-Augmented Generation (RAG) using Flink and Kafka, along with clustering techniques to improve chunking in RAG systems. Dive in to see how these technologies are shaping the future!

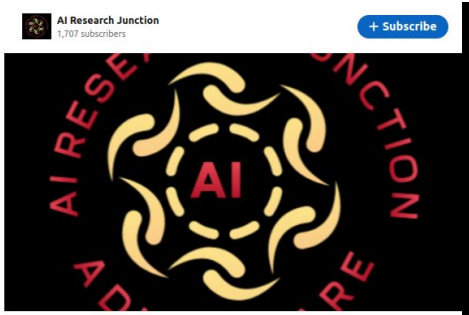
Generative AI

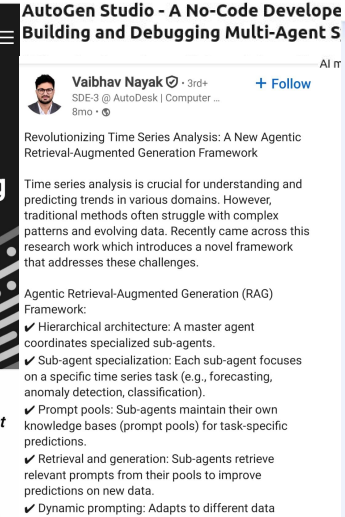
- Agentic-RAG: A Hierarchical Multi-Agent Framework for Enhanced Time Series Analysis. In this post, Sana Hassan explores a novel framework called Agentic-RAG, developed by researchers from IIT Dharwad and TCS Research, for enhanced time series analysis. The framework tackles common challenges like high dimensionality and distribution shifts by employing a hierarchical, multi-agent architecture. Each sub-agent, fine-tuned for specific forecasting or anomaly detection tasks, uses pre-trained language models (SLMs) and retrieves prompts from a specialized knowledge pool. This dynamic system results in more accurate and flexible predictions, significantly outperforming traditional methods across diverse datasets. This post highlights an essential advancement in time series analysis, emphasizing the integration of SLMs with a two-tiered attention mechanism. The researchers' use of Direct Preference Optimization (DPO) and parameter-efficient fine-tuning techniques enhances the performance of SLMs, allowing them to better handle long-range dependencies and adapt to evolving data patterns. The approach demonstrates clear superiority across key datasets, showcasing how SLMs can transcend their traditional limitations in text analysis and successfully apply to complex time series tasks.
- Inside GameNGen: Google DeepMind's New Model that can Simulate Entire 1993's DOOM Game in Real Time. This post by Jesus Rodriguez delves into Google DeepMind's GameNGen, a groundbreaking diffusion model capable of simulating the entire 1993 game DOOM in real time. Unlike traditional game engines, GameNGen relies on neural networks to handle complex tasks such as tracking health, ammo, and player interactions, maintaining the game's immersive experience. Built on an enhanced version of Stable Diffusion v1.4, the model demonstrates that AI can effectively simulate video games in real time, showcasing significant progress in generative AI applications for gaming. What stands out in this post is the innovative training process behind GameNGen. Rather than relying on human gameplay data, which is challenging to scale, an automated agent is trained to interact with the game environment, collecting data for the generative model. This efficient data collection allows the model to simulate long sequences with high fidelity. GameNGen's ability to blur the line between neural simulations and traditional game engines points to an exciting future where games may be auto-generated using AI.
- Table-Augmented Generation (TAG): A Unified Approach for Enhancing Natural Language Querying over Databases. This post introduces Table-Augmented Generation (TAG), a unified framework that UC Berkeley and Stanford University researchers

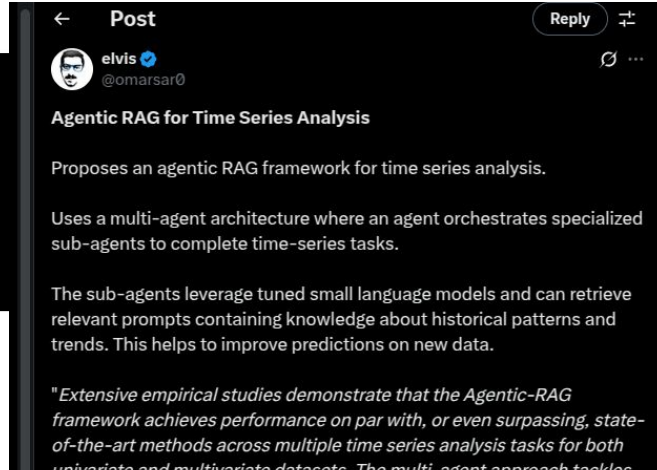


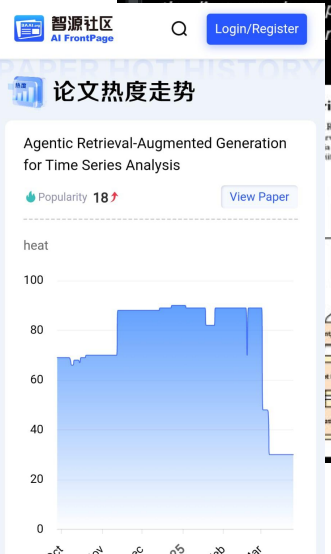
Decoding the language of time: AI that understands time series as seamlessly as text

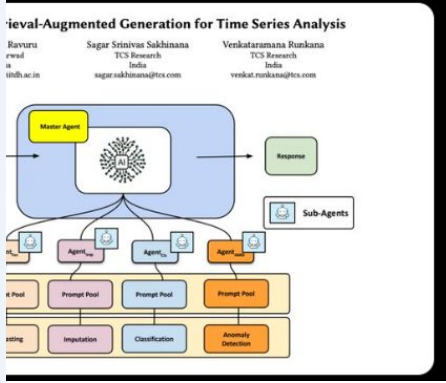
By Sina Pakazad, Vice President, Data











collaboration [16], enabling them to manage dynamic workflows and tackle intricate problems.

The convergence of RAG and agentic intelligence has given rise to Agentic Retrieval-Augmented Generation (Agen
RAG) [17], a paradigm that integrates agents into the RAG pipeline. Agentic RAG enables dynamic retrieval strateg
contextual understanding, and iterative refinement [18], allowing for adaptive and efficient information processi
Unlike traditional RAG, Agentic RAG employs autonomous agents to orchestrate retrieval, filter relevant informati
and refine responses, excelling in scenarios requiring precision and adaptability.

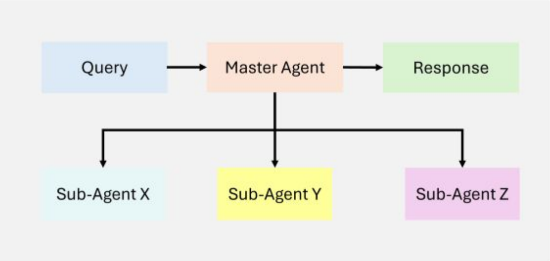


Figure 13: An illustration of Hierarchical Agentic RAG

- **Scalability**: Distributing tasks across multiple agent tiers enables handling of highly complex or multi-faceted queries.
- **Enhanced Decision-Making**: Higher-level agents apply strategic oversight to improve overall accuracy and coherence of responses.

Challenges

- **Coordination Complexity**: Maintaining robust inter-agent communication across multiple levels can increase orchestration overhead.
- **Resource Allocation**: Efficiently distributing tasks among tiers to avoid bottlenecks is non-trivial.

Use Case: Financial Analysis System

Prompt: What are the best investment options given the current market trends in renewable energy?

System Process (Hierarchical Agentic Workflow):

1. **Top-Tier Agent**: Assesses the query's complexity and prioritizes reliable financial databases and economic indicators over less validated data sources.
2. **Mid-Level Agent**: Retrieves real-time market data (e.g., stock prices, sector performance) from proprietary APIs and structured SQL databases.
3. **Lower-Level Agent(s)**: Conducts web searches for recent pc mendation systems that track expert opinions and news analy
4. **Aggregation and Synthesis**: The top-tier agent compiles the policy insights.

Response:
Integrated Response: "Based on current market data, renewable energ the past quarter, driven by supportive government policies and height that wind and solar sectors, in particular, may experience continued n like green hydrogen present moderate risk but potentially high returns

README

References

Research Papers on Agentic RAG

1. Single-Agent RAG (Router-Based)

- Search-o1: Agentic Search-Enhanced Large Reasoning Models https://arxiv.org/abs/2501.05366

2. Multi-Agent Agentic RAG

- Agentic Retrieval-Augmented Generation for Time Series Analysis https://arxiv.org/abs/2408.14484

3. Corrective Agentic RAG

- Agentic AI-Driven Technical Troubleshooting for Enterprise Systems https://arxiv.org/abs/2412.12006
- Corrective RAG (CRAG) https://langchain-ai.github.io/langgraph/tutorials/rag/langgraph_crag/
- Corrective Retrieval Augmented Generation https://arxiv.org/abs/2401.15884
- Agentic AI-Driven Technical Troubleshooting for Enterprise Systems https://arxiv.org/abs/2412.12006

4. Adaptive Agentic RAG

- Langgraph Adaptive RAG https://langchain-ai.github.io/langgraph/tutorials/rag/langgraph_adaptive_rag/
- MBA-RAG: A Bandit Approach for Adaptive Retrieval-Augmented https://arxiv.org/abs/2412.01572
- CtrlA: Adaptive Retrieval-Augmented Generation via Inherent Control https://arxiv.org/abs/2405.18727
- Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity https://arxiv.org/abs/2403.14403
- AT-RAG: An Adaptive RAG Model Enhancing Query Efficiency with Topic Filtering and Iterative Reasoning https://arxiv.org/abs/2410.12886

5. Graph-Based Agentic RAG

- GeAR: Graph-enhanced Agent for Retrieval-augmented Generation https://arxiv.org/abs/2412.18431
- Agent-G: An Agentic Framework for Graph Retrieval Augmented Generation https://openreview.net/forum?id=g2C947jjjQ

How to Cite

4.3 Hierarchical Agentic RAG Systems

Hierarchical Agentic RAG: [17] systems employ a structured, multi-tiered approach to information retrieval and processing, enhancing both efficiency and strategic decision-making as shown in Figure 13. Agents are organized in a hierarchy, with higher-level agents overseeing and directing lower-level agents. This structure enables multi-level decision-making, ensuring that queries are handled by the most appropriate resources.

Workflow

1. **Query Reception**: A user submits a query, received by a *top-tier agent* responsible for initial assessment and delegation.
2. **Strategic Decision-Making**: The top-tier agent evaluates the query's complexity and decides which subordinate agents or data sources to prioritize. Certain databases, APIs, or retrieval tools may be deemed more reliable or relevant based on the query's domain.
3. **Delegation to Subordinate Agents**: The top-tier agent assigns tasks to lower-level agents specialized in particular retrieval methods (e.g., SQL databases, web search, or proprietary systems). These agents execute their assigned tasks independently.
4. **Aggregation and Synthesis**: The results from subordinate agents are collected and integrated by the higher-level agent, which synthesizes the information into a coherent response.
5. **Response Delivery**: The final, synthesized answer is returned to the user, ensuring that the response is both comprehensive and contextually relevant.

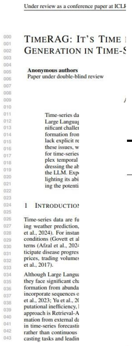
Key Features and Advantages.

- **Strategic Prioritization**: Top-tier agents can prioritize data sources or tasks based on query complexity,


Towards Resolution-Aware Retrieval Augmented Zero-Shot Forecasting

Abstract

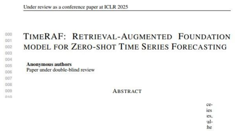
1 Introduction


Under review as a conference paper at ICLR

TIMERAG: IT'S TIME
GENERATION IN TIME-S

Anonymous authors
Paper under double-blind review

1 INTRODUCTION


Under review as a conference paper at ICLR 2025

TIMERAF: RETRIEVAL-AUGMENTED FOUNDATION MODEL FOR ZERO-SHOT TIME SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT


Retrieval-Augmented Diffusion Models for Time Series Forecasting

Abstract

1 Introduction

HiPerRAG: High-Performance Retrieval Augmented Generation for Scientific Insights

Ozan Gokdemir[1,2†*], Carlo Siebenschuh[1,2†], Alexander Brace[1,2†], Azton Wells[1†], Brian Hsu[1,2†], Kyle Hippe[1,2], Priyanka V. Setty[1,2], Aswathy Ajith[2], J. Gregory Pauloski[2], Varuni Sastry[1], Sam Foreman[1], Huihuo Zheng[1], Heng Ma[1], Bharat Kale[1], Nicholas Chia[1], Thomas Gibbs[3], Michael E. Papka[1,4], Thomas Brettin[1], Francis J. Alexander[1], Anima Anandkumar[5], Ian Foster[1,2], Rick Stevens[1,2], Venkatram Vishwanath[1], Arvind Ramanathan[1,2],
[1]Argonne National Laboratory, Lemont, Illinois, USA [2]The University of Chicago, Chicago, Illinois, USA
[3]NVIDIA Inc., Santa Clara, California, USA [4]University of Illinois Chicago, Chicago, Illinois, USA
[5]California Institute of Technology, Pasadena, California, USA
†Joint first authors *Corresponding authors: ramanathana@anl.gov, stevens@anl.gov, venkat@anl.gov

bstract
volume of scientific literature is growing exponentially, lead- to underutilized discoveries, duplicated efforts, and limited ss-disciplinary collaboration. Retrieval-Augmented Generation G) offers a way to assist scientists by improving the factual- of Large Language Models (LLMs) in processing this influx of rmation. However, scaling RAG to handle millions of articles oduces significant challenges, including the high computational ts associated with parsing documents and embedding scientific wledge, as well as the algorithmic complexity of aligning these resentations with the nuanced semantics of scientific content. address these issues, we introduce HiPerRAG, a RAG workflow powered by high performance computing (HPC) to index and retrieve knowledge from more than 3.6 million scientific articles. At its core are Oreo, a high-throughput model for multimodal document parsing, and ColTrast, a query-aware encoder fine-tuning algorithm that enhances retrieval accuracy by using contrastive

Thomas Brettin[1], Francis J. Alexander[1], Anima Anandkumar[5], Ian Foster[1,2], Rick Stevens[1,2], Venkatram Vishwanath[1], Arvind Ramanathan[1,2], . 2025. HiPerRAG: High-Performance Retrieval Augmented Generation for Scientific Insights. In *Platform for Advanced Scientific Computing Conference (PASC '25), June 16–18, 2025, Brugg-Windisch, Switzerland.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3732775.3733586

1 Introduction

Over the past century, the volume of scientific publications has increased rapidly. Today, this trend continues at an unprecedented pace. For example, the National Science Foundation reported a 50-fold increase in open-access scientific publications between 2013 and 2022 [22]. In the biomedical field alone, PubMed processed approximately 1.69 million articles last year, averaging more than three articles per minute [45]. Despite this vast output, researchers' capacity to keep up is limited, estimated at 22 articles per month [74]. This overload of information has significant consequences


Yeah. Good morning, everyone today. I'll be speaking about parameter efficient.


ICML
International Conference
On Machine Learning
PARAMETER-EFFICIENT QUANTIZED MOE MEETS VISION-LANGUAGE INSTRUCTION TUNING FOR SEMICONDUCTOR ELECTRON MICROGRAPH ANALYSIS
Venkataramana Runkana

TATA INNOVATION CENTER
S. Sakhinana et al. (ICML 2024)
AI for Semiconductor Manufacturing
ICML 2024


ICLR
International Conference On
Learning Representations


IJCAI/2023
The 32nd International
Joint Conference
on Artificial Intelligence
19th-25th
August 2023
Sheraton Grand Macao,
Macao SAR


PREVIEW ONLY, NOT LIVE STREAMED!


ICLR
International Conference On
Learning Representations