subatomic

Evaluation Scope

Project Name	Task Name	Assigned To
SUBATOMIC ATLAS	Support CSV and Various File Types in Core and Sales Agents	Aaron
SUBATOMIC ATLAS	Implement HITL with Dynamic Scope for Core and Sales Agents	Aaron
SUBATOMIC ATLAS	Enable Multi-dimensional Wiki Management (Role-based)	Aaron
SUBATOMIC ATLAS	Navigation Support on Streaming Pages	Emmanuel
SUBATOMIC ATLAS	Validate Global Styling Across All Pages	Emmanuel
SUBATOMIC ATLAS	Improve Memory Report and Management (Episodic, Procedural, Semantic Types)	Emmanuel
SUBATOMIC ATLAS	Generate Standard Documentation Format for RAG Ingestion & Retrieval	Emmanuel
SUBATOMIC ATLAS	Implement RBAC Across All Existing Modules	Emmanuel
SUBATOMIC ATLAS	Regression Testing for RAG Retrieval (Aligned With Supervised Evaluation)	Emmanuel
SUBATOMIC ATLAS	Document Plan for Scaling Vector Store During Document Ingestion	Emmanuel
SUBATOMIC ATLAS	Benchmark & Reduce Latency in Existing Agents	Emmanuel
SUBATOMIC ATLAS	Reduce User Session Time and Implement MFA	Emmanuel
SUBATOMIC ATLAS	Migrate All Agents to Deep Agents Module (LangGraph)	Emmanuel
SUBATOMIC ATLAS	Dynamic Connection to Notification System for Scheduled File Ingestions	Emmanuel
REINFORCEMENT LEARNING FOR CHAT WITH ATLAS	Configure and Store RLHF Dataset	Luis
REINFORCEMENT LEARNING FOR CHAT WITH ATLAS	Implement Atlas Agent Insights Analyzer	Luis
REINFORCEMENT LEARNING FOR CHAT WITH ATLAS	Pattern Recognition Agent Based on Integrated Data Source	Luis
SUBATOMIC NEXUS	Integrate Full Workflow in UI	Aaron
SUBATOMIC NUCLEUS	Integrate Full Workflow in UI	Aaron
SUBATOMIC NUCLEUS	Scale and Improve AI Co-Worker Tool Generation with Dynamic Tool Calling	Aaron
MULTI-DIMENSIONAL CONTRACT COMPARISON REVIEWER	Add Agent for Client Qualification ("Less Restrictive" Terms & Conditions)	Christopher
MULTI-DIMENSIONAL CONTRACT COMPARISON REVIEWER	Implement Deep Agent (LangChain) with GPT-4.1 Orchestration and Planning	Christopher
VANTAGE FINANCIAL	Deploy Automatic DAGs (Airflow) to Azure	Emmanuel
VANTAGE FINANCIAL	Client-Specific Memories in Agenda Compilation	Aaron
VANTAGE FINANCIAL	Enhanced HITL for Visual Agendas	Aaron
VANTAGE FINANCIAL	Scheduled & On-Demand Agenda Compilation	Aaron

Task Evaluation Definition

Task Name: Support CSV and Various File Types in Core and Sales Agents

Evaluation Aspect	Definition
Expected Result / Outcome	Core and Sales Agents that can successfully ingest, parse, and process CSV, Excel, PDF, and DOCX files with correct mapping, error handling, and validation.
Success Metrics	End-to-end ingestion tests pass for all supported types.
Measurement Method	Automated tests; ability for user to retrieve and utilize ingested knowledge.
Quality Standards	Correct mapping, error handling, validation; use of proven libraries, log errors clearly, cover edge cases, maintain schema mapping standards.
Acceptance Criteria	End-to-end ingestion tests pass for all supported types.
Review / Validation Owner	Not specified in context
Constraints	Baseline ingestion pipeline must be operational; coordinate with Emmanuel on notification system for ingestion.
Notes	Coordinate testing with Emmanuel for notification triggers and system integration. Prepare for future expansion to additional formats.

Task Name: Implement HITL with Dynamic Scope for Core and Sales Agents

Evaluation Aspect	Definition
Expected Result / Outcome	Configurable HITL system with: review queue/UI, rules for scope assignment, integration in agent action pipeline; documentation provided.
Success Metrics	Test scenarios validate scope adjustment and review flow.
Measurement Method	Testing of agent-HITL integration with E2E scenarios; positive user feedback on review experience.
Quality Standards	Middleware modularity, adjustable configuration, seamless UX for reviewers.
Acceptance Criteria	HITL can be enabled/disabled; scope adjusts dynamically; positive user feedback on review experience.
Review / Validation Owner	Not specified in context
Constraints	Coordination across backend, frontend, and integration with existing flows.
Notes	Critical for regulated workflows; should tie into compliance reporting where applicable.

Task Name: Enable Multi-dimensional Wiki Management (Role-based)

Evaluation Aspect	Definition
Expected Result / Outcome	Wiki system supporting multiple "dimensions" (roles/departments), with permissions and structured navigation.
Success Metrics	Passes positive and negative access tests.
Measurement Method	Testing for permission leaks; QA/users test access/edit capabilities.
Quality Standards	Fine-grained permissions, clear role assignment, logged permission checks.
Acceptance Criteria	Only authorized users can access/edit appropriate sections; tested by QA/users.
Review / Validation Owner	Not specified in context
Constraints	Depends on RBAC implementation, current Wiki infrastructure.
Notes	Considerability for future expansion (e.g., tags, departments, projects).

Task Name: Navigation Support on Streaming Pages

Evaluation Aspect	Definition
Expected Result / Outcome	Streaming page navigation works smoothly—no dropped streams, correct browser/app history updates, passes usability tests.
Success Metrics	All navigation operations work without dropped data or errors.
Measurement Method	Testing with long-running streams; usability testing.
Quality Standards	Decouple stream state from route, user informed of state changes.
Acceptance Criteria	All navigation operations work without dropped data or errors.
Review / Validation Owner	Not specified in context
Constraints	Medium complexity due to streaming state management.
Notes	Ensure accessibility for all users/devices.

Task Name: Validate Global Styling Across All Pages

Evaluation Aspect	Definition
Expected Result / Outcome	No visual inconsistencies remain; all pages adhere to style guide. Acceptance: UI review checklist passes.
Success Metrics	No reported styling inconsistencies after release.
Measurement Method	Automated scan (Storybook/Chromatic), manual visual review, UI review checklist.
Quality Standards	Global styles per style guide, centralized styling, no inline overrides.
Acceptance Criteria	UI review checklist passes; no reported styling inconsistencies after release.
Review / Validation Owner	Not specified in context
Constraints	Browser compatibility, legacy components.
Notes	Consider accessibility (WCAG) compliance during review.

Task Name: Improve Memory Report and Management (Episodic, Procedural, Semantic Types)

Evaluation Aspect	Definition
Expected Result / Outcome	Memory subsystem with clear distinctions and reporting for all types, updated management UI/API, documented usage.
Success Metrics	All memories properly classified and reported; acceptance tests pass for examples of each type.
Measurement Method	Validation by querying/testing examples of each memory type.
Quality Standards	Document memory type criteria; log classification for traceability.

Evaluation Aspect	Definition
Acceptance Criteria	Acceptance tests pass for examples of each type; all memories classified and reported.
Review / Validation Owner	Not specified in context
Constraints	Medium complexity; requires design and data migration.
Notes	Lay groundwork for memory-based agent improvements.

Task Name: Generate Standard Documentation Format for RAG Ingestion & Retrieval

Evaluation Aspect	Definition
Expected Result / Outcome	Specified documentation format (template + schema), validator/converter tool, docs for users/authors; passes ingestion/retrieval tests.
Success Metrics	100% ingestion conformance; no ingestion errors; users can convert legacy docs.
Measurement Method	Ingestion/retrieval tests; validation/conformance checks.
Quality Standards	Version schema; CI check for docs pre-ingestion; example templates, changelog for spec evolution.
Acceptance Criteria	100% ingestion conformance; no ingestion errors; users can convert legacy docs.
Review / Validation Owner	Not specified in context
Constraints	Medium complexity; depends on content diversity.
Notes	Include example templates, changelog for spec evolution.

Task Name: Implement RBAC Across All Existing Modules

Evaluation Aspect	Definition
Expected Result / Outcome	Comprehensive RBAC implementation, passes penetration and misuse tests; documentation for roles/permissions; admin UI.
Success Metrics	Zero unauthorized access in security testing; users have appropriate access only.
Measurement Method	Penetration testing; misuse testing; access control validation.
Quality Standards	Test with least-privilege; log permission failures; document all permission rules.
Acceptance Criteria	Zero unauthorized access in security testing; users have appropriate access only.
Review / Validation Owner	Not specified in context
Constraints	Wide scope and risk if permissiveness too high.
Notes	Prioritize critical modules; coordinate with Aaron for agent-integrated RBAC.

Task Name: Regression Testing for RAG Retrieval (Aligned With Supervised Evaluation)

Evaluation Aspect	Definition
Expected Result / Outcome	Regression test suite, test result reports, bug tickets for issues; reproducible and automated.
Success Metrics	All regression tests pass after new changes; test suite covers >90% of typical scenarios.
Measurement Method	Automated testing; test result reports; bug tracking.
Quality Standards	Version test cases; automate test runs; peer review assertions and coverage.
Acceptance Criteria	All regression tests pass after new changes; >90% scenario coverage.
Review / Validation Owner	Validation with Aaron.
Constraints	Medium effort; scales with scope/coverage of test suite.
Notes	Coordinate test plan with Aaron to capture AI-specific nuances.

Task Name: Document Plan for Scaling Vector Store During Document Ingestion

Evaluation Aspect	Definition
Expected Result / Outcome	Documented scalability plan (architecture, migration, risks), review with team; update backlog based on recommendations.
Success Metrics	Plan approved, risk/impact understood, ready for implementation as needed.
Measurement Method	Team review and approval.
Quality Standards	Modular, vendor-agnostic design.
Acceptance Criteria	Plan approved; risk/impact understood.
Review / Validation Owner	Team (specific owner not stated).
Constraints	Medium complexity (mostly design); must consider data migration downtime and vendor limitations.
Notes	Consider growth projections and multi-region design if needed.

Task Name: Benchmark & Reduce Latency in Existing Agents

Evaluation Aspect	Definition
Expected Result / Outcome	Benchmark report, optimized agent code, measurable reduction in response times; supporting documentation.
Success Metrics	Quantitative reduction in average/max response time; stable ops post-deployment.
Measurement Method	Before/after performance benchmarks; response time logging and validation.
Quality Standards	Isolate optimizations for validation; track pre/post metrics for evidence.
Acceptance Criteria	Quantitative reduction in average/max response time; stable ops post-deployment.
Review / Validation Owner	Not specified in context
Constraints	Some analysis, some code refactor; risk of breaking downstream logic.
Notes	Share learnings with broader team for similar optimizations.

Task Name: Reduce User Session Time and Implement MFA

Evaluation Aspect	Definition
Expected Result / Outcome	System with reduced session time and required MFA for all users; passes penetration and usability tests.
Success Metrics	100% MFA enforcement and shorter session times on all user logins; zero bypasses.
Measurement Method	Penetration testing, usability testing, validation of session timeout and MFA flow.
Quality Standards	Provide fallback for lost MFA devices, audit/log all auth attempts.
Acceptance Criteria	100% MFA enforcement and shorter session times on all user logins; zero bypasses.
Review / Validation Owner	Not specified in context
Constraints	MFA provider downtime; user friction; mobile auth support.
Notes	Communicate change to users in advance, prepare support resources.

Task Name: Migrate All Agents to Deep Agents Module (LangGraph)

Evaluation Aspect	Definition
Expected Result / Outcome	All agents operational via Deep Agents, passes regression/stability tests; legacy code retired.
Success Metrics	100% of agents running via Deep Agents; equal or better stability/performance.
Measurement Method	Regression/stability testing, before/after benchmarks, documentation checks.
Quality Standards	Test in isolation, maintain rollout plan for rollback if issues emerge.
Acceptance Criteria	100% of agents running via Deep Agents; equal or better stability/performance.
Review / Validation Owner	Not specified in context
Constraints	Agent-specific logic may make migration nontrivial; complex task.
Notes	Identify and prioritize high-impact agents first.

Task Name: Dynamic Connection to Notification System for Scheduled File Ingestions

Evaluation Aspect	Definition
Expected Result / Outcome	Notifications (email, in-app, etc.) triggered correctly for scheduled and ad hoc ingestion events; configuration options for notification recipients.
Success Metrics	Timely, accurate notifications for all ingestion schedules; no excess/duplicate alerts.
Measurement Method	Testing events in sandbox; review of notification logs/events.
Quality Standards	Debounce/throttle notifications; flexible recipient targeting.
Acceptance Criteria	Timely, accurate notifications for all ingestion schedules; no excess/duplicate alerts.
Review / Validation Owner	Not specified in context
Constraints	Race conditions (duplicate/truncated events), noisy notifications, access control.
Notes	Initial rollout with key teams, expand recipients as needed.

Task Name: Configure and Store RLHF Dataset

Evaluation Aspect	Definition
Expected Result / Outcome	Secure data store (e.g., S3, Databricks, GCP Bucket) with role-based access, versioning, clear documentation of data schema and use protocols.
Success Metrics	Dataset can be safely accessed/updated by intended users; audit/logs confirm security.
Measurement Method	Test data access by roles; audit/log reviews.
Quality Standards	Encrypt at rest/in transit, automate audit logs, restrict data egress.

Evaluation Aspect	Definition
Acceptance Criteria	Dataset can be safely accessed/updated by intended users; audit/logs confirm security.
Review / Validation Owner	Permissions validated with Aaron before go-live.
Constraints	Security misconfigurations, data versioning errors, cost overage.
Notes	Periodic backup script; roles and permissions need validation.

Task Name: Implement Atlas Agent Insights Analyzer

Evaluation Aspect	Definition
Expected Result / Outcome	Analytics dashboard/report comparing conversation quality, user satisfaction, and agent performance before and after RLHF application.
Success Metrics	Clear, actionable report produced; team agrees on interpretation of results.
Measurement Method	Report/dashboard creation and review; team interpretation process.
Quality Standards	Use blinded review; verify statistical significance where feasible.
Acceptance Criteria	Clear, actionable report produced; team agreement on interpretation.
Review / Validation Owner	Aaron (review methodology for technical validity).
Constraints	Data integrity; ambiguous metrics; insufficient data split for A/B.
Notes	Plan ongoing analysis cadence.

Task Name: Pattern Recognition Agent Based on Integrated Data Source

Evaluation Aspect	Definition
Expected Result / Outcome	Pattern recognition agent, code repo, sample analysis results, test suite signed off by Aaron.
Success Metrics	Patterns match expectations, useful signals delivered; reviewed and approved by Aaron.
Measurement Method	Testing and documentation; analysis/interpretation of output; Aaron signs off on evaluation.
Quality Standards	Build incremental, start with simple patterns, validate results with real-world feedback.
Acceptance Criteria	Patterns match expectations; useful signals delivered; reviewed and approved by Aaron.
Review / Validation Owner	Aaron
Constraints	Data volume/quality, false positives, model overfitting/generalizability.
Notes	Use explainability tools as feasible.

Task Name: Integrate Full Workflow in UI (Subatomic Nexus)

Evaluation Aspect	Definition
Expected Result / Outcome	Intuitive UI supporting all workflow stages, tested with "happy" and edge-case paths, demo-ready.
Success Metrics	Users can complete full workflow in UI; feedback indicates clarity and usability.
Measurement Method	User testing; feedback collection; readiness of demo.
Quality Standards	Use stepper/progress indicator; surface clear help/errors; test all workflow branches.
Acceptance Criteria	Users can complete full workflow in UI; positive feedback on clarity and usability.
Review / Validation Owner	Not specified in context
Constraints	Workflow step consistency, error propagation/display, real-time status sync.
Notes	Reuse UI components where possible for maintainability.

Task Name: Integrate Full Workflow in UI (Subatomic Nucleus)

Evaluation Aspect	Definition
Expected Result / Outcome	UI supports all Nucleus workflow steps with guidance, error handling, permissions integration. Beta users can form AI teams successfully.
Success Metrics	End-users can successfully create/deploy AI teams; workflow tracked/logged for support.
Measurement Method	Beta feedback/testing; workflow tracking and logging.
Quality Standards	Inline progress, save/resume draft, confirm each step.
Acceptance Criteria	End-users can successfully create/deploy AI teams; workflow is tracked/logged.
Review / Validation Owner	Not specified in context
Constraints	Handling workflow exceptions; tool assignment edge cases.

Evaluation Aspect	Definition
Notes	Sync workflow logic with backend to minimize translation bugs.

Task Name: Scale and Improve AI Co-Worker Tool Generation with Dynamic Tool Calling Agentic Pattern

Evaluation Aspect	Definition
Expected Result / Outcome	Nucleus supports agentic tool calling: dynamic selection/invocation, logging, and fallback; tests confirm correct tool assignment/ execution.
Success Metrics	Dynamic tool assignment works seamlessly; demonstrated with real user workflows.
Measurement Method	Testing with multiple team configurations; demonstration with user workflows.
Quality Standards	Modular agent logic; clear fallback/exception flows; documented invariant behaviors.
Acceptance Criteria	Dynamic tool assignment works seamlessly; tests confirm correct tool assignment/execution.
Review / Validation Owner	Not specified in context
Constraints	Tool permissioning, model ambiguity, error handling in tool failures.
Notes	Coordinate tool metadata schemas with Emmanuel if shared with other modules.

Task Name: Add Agent for Client Qualification ("Less Restrictive" Terms & Conditions)

Evaluation Aspect	Definition
Expected Result / Outcome	Agent code fully integrated, documented, and validated with real-world contract samples; accuracy metrics reported.
Success Metrics	Agent correctly flags clients per criteria; passes legal/user evaluation.
Measurement Method	Validation outputs/test cases; legal/user evaluation; reporting of accuracy metrics.
Quality Standards	Collaborate with legal SME, build explainable output, unit test extensively with edge cases.
Acceptance Criteria	Agent correctly flags clients per criteria; passes legal/user evaluation.
Review / Validation Owner	SME/legal for feedback loop (no explicit owner for final acceptance).
Constraints	Ambiguous contract language, model calibration, defining/agreement of "less restrictive".
Notes	Feedback loop with SME/legal required for continuous improvement.

Task Name: Implement Deep Agent (LangChain) with GPT-4.1 Orchestration and Planning Middleware

Evaluation Aspect	Definition
Expected Result / Outcome	Orchestrator runs all sub-agents per plan, logs all steps, user can view/modify execution plan.
Success Metrics	All agents orchestrated successfully in stepwise fashion; execution is traceable and auditable.
Measurement Method	Testing with complex cases; plan exposure/edit in UI; log review.
Quality Standards	Log all plan steps, provide fallback in error states, modular for future sub-agent expansion.
Acceptance Criteria	All agents orchestrated successfully in stepwise fashion; execution is traceable and auditable.
Review / Validation Owner	Not specified in context
Constraints	Step consistency, sub-agent misalignment, prompt design.
Notes	Coordinate with backend team for integration; document for future handoffs.

Task Name: Deploy Automatic DAGs (Airflow) to Azure

Evaluation Aspect	Definition
Expected Result / Outcome	DAGs successfully deployed and running in Azure; documented CI/CD pipeline, monitoring alerts in place.
Success Metrics	100% automated deploy/redeploy; no missing/failed DAGs post-pipeline run.
Measurement Method	Test deploy pipeline; health/monitoring logs.
Quality Standards	Immutable infra where possible, alert on pipeline failure, periodic validation jobs.
Acceptance Criteria	100% automated deploy/redeploy; no missing/failed DAGs post-pipeline run.
Review / Validation Owner	Not specified in context
Constraints	Azure perm/access issues, DAG cross-dependencies, monitoring/alert noise.
Notes	Schedule routine audits for pipeline and DAG failures.

Task Name: Client-Specific Memories in Agenda Compilation

Evaluation Aspect	Definition
Expected Result / Outcome	Agenda compiler accesses/utilizes client memories; user sees accurate, tailored output in end-to-end testing.

Evaluation Aspect	Definition
Success Metrics	Agendas produced are more accurate/relevant for specific clients; passes user evaluation.
Measurement Method	Testing with simulated/real clients; user-facing accuracy validation.
Quality Standards	Modularize for future clients; log memory access for debugging.
Acceptance Criteria	Agendas more accurate/relevant for specific clients; passes user evaluation.
Review / Validation Owner	Not specified in context
Constraints	Memory schema mismatches, permission edge cases, interface stability.
Notes	Coordinate memory schema unification with Emmanuel if applicable.

Task Name: Enhanced HITL for Visual Agendas

Evaluation Aspect	Definition
Expected Result / Outcome	Visual agenda HITL system implemented, with clear review/approval UI, audit logs, and configuration options.
Success Metrics	Users can reliably review/approve agendas; logs capture all actions; no unapproved outputs.
Measurement Method	End-to-end testing; user training/docs; audit log review.
Quality Standards	Modular reviewer UI, reviewer delegation/escalation, automate notifications.
Acceptance Criteria	Users can reliably review/approve agendas; logs capture all actions; no unapproved outputs.
Review / Validation Owner	Not specified in context
Constraints	Reviewer availability; feature creep; audit log completeness.
Notes	Consider regulatory requirements for audit trail retention.

Task Name: Scheduled & On-Demand Agenda Compilation

Evaluation Aspect	Definition
Expected Result / Outcome	System can create agendas per schedule, or when users press “compile”, with process tracking and usable error messages.
Success Metrics	Agendas compile as expected in both modes, no missed/double runs, status clear to users.
Measurement Method	QA for schedule edge cases; monitoring and error logging.
Quality Standards	Debounce jobs where needed, track all runs for auditing.
Acceptance Criteria	Agendas compile as expected in both modes, status clear to users, no missed/double runs.
Review / Validation Owner	Not specified in context
Constraints	Job queue reliability, duplicate runs, time zone support.
Notes	Schedule periodic revalidation of trigger/schedule reliability.