S1

ID	Task	DoD	Estimate (h)	Owner	Dependencias	Riesgo	Criterios de aceptación
S1-01	Inicializar repo con `uv` y estructura de carpetas	`pyproject.toml`, `uv.lock`, árbol base creado; `.gitignore` para `logs_local/`	6	This	—	B	`uv sync` funciona; estructura `/src`, `/configs`, `/logs`, `/analytics/parquet`, `/tests`, `/reports` creada
S1-02	Configuración CI (lint/tests) + pre-commit	Workflow GH Actions (lint+tests) pasa; hooks activos	2	This	S1-01	B	PR con checks verdes; `pre-commit run -a` limpio
S1-03	Loader GSM8K + normalización	Función `load_batch(n, seed)` y normalización de respuestas numéricas	6	José	S1-01	M	Lee ≥200 ítems; pruebas unitarias básicas pasan
S1-04	Evaluador exact-match (GSM8K)	`evaluate_exact_match(y_true, y_pred)` con normalización	4	José	S1-03	B	Test unitario con casos borde (comas, espacios)
S1-05	Contador de tokens (prompt+completion)	`count_tokens(event)` integrado al logger	3	This	S1-01	B	Test unitario simple; devuelve enteros consistentes
S1-06	Baseline runner ≥200 (single-call)	Flow Prefect que ejecuta baseline con logging JSONL	8	This	S1-03, S1-04, S1-05, S1-09	M	Artefactos: `/logs/events/*.jsonl`, `/analytics/parquet/baseline.parquet`
S1-07	Núcleo T-A-S (k=1) con control de temperatura	Módulos `thesis()`, `antithesis()`, `synthesis()`; T tesis=0.7, T síntesis=0.2	10	Julio	S1-01	M	Funciones puras probadas con mocks; parámetros en `.env`/`configs/`
S1-08	Prefect flow T-A-S (orquestación)	Tareas encadenadas T→A→S (k=1) con manejo de errores	6	This	S1-07	M	`prefect deploy` o script `python -m src.flows.tas` corre en seco
S1-09	Logger JSONL + sanitización/anonimización	`log_event_jsonl(record)` usa `sanitize()` y hashing con sal	6	Lorena	S1-01	M	CoT nunca sale de `logs_local/`; JSONL compartido solo con resúmenes
S1-10	Agregación Parquet (analytics)	Job que convierte JSONL→Parquet por run	4	José	S1-06, S1-09	B	Archivo `/analytics/parquet/*.parquet` legible con pandas/pyarrow
S1-11	Unit tests mínimos (parser/evaluator/sanitize/flow)	≥4 tests críticos; cobertura básica	8	Julio (4) / José (3) / Lorena (1)	S1-03..S1-09	M	`pytest -q` OK en CI
S1-12	Pilot run T-A-S (~50)	Ejecución real; logs locales y eventos compartidos generados	4	Julio	S1-07..S1-11	M	`/logs_local/*` con CoT; `/logs/events/*` sin CoT; sin fallos
S1-13	McNemar + KPIs (baseline vs T-A-S)	Script estadístico; tabla de métricas; cálculo ΔAcc y tokens	5	José	S1-06, S1-12, S1-10	M	Reporte tabular con p-valor, ΔAcc, costo tokens
S1-14	README (cómo correr y reproducir)	Secciones: setup `uv`, baseline, T-A-S, métricas	4	Valeria	S1-01..S1-13	B	Seguir README reproduce baseline y piloto T-A-S
S1-15	Reporte corto Sprint1.md	Resumen de resultados, riesgos, próximos pasos	3	Valeria	S1-13	B	`/reports/Sprint1.md` con KPIs y conclusiones
S1-16	Templating de prompts + hashing de prompt/resp	Funciones utilitarias y pruebas	3	Julio	S1-01	B	`prompt_hash`, `response_hash` presentes en logs

S2

ID	Task	DoD	Est. (h)	Owner	Dependencias	Riesgo	Criterios de aceptación
S2-01	Escalado Prefect + retries/backoff + rate-limit aware	Flow con backoff (1s/2s/4s+jitter) y manejo de límites	6	This	S1 infra	M	Runs no fallan por rate limit; logs de reintentos en JSONL
S2-02	Parametrizar jitter de temperaturas y seeds	Config central: Tesis {0.65,0.70,0.75}; k=1; seeds {101,202,303}	3	Julio	S1 T-A-S	B	`configs/model.yaml` actualizado; tests de lectura de config
S2-03	Implementar MAMV (3 instancias)	Síntesis por mayoría simple con trazabilidad por instancia	10	Julio	S2-02	M	`synthesis_mamv()` retorna decisión y votos; unit test básico
S2-04	Reutilizar los mismos 200 ítems de S1 (seed=42, estratos)	Runner toma exactamente el mismo set (emparejamiento 1-a-1)	3	José	S1 loader	B	Hash de IDs coincide con baseline; dump de IDs versionado
S2-05	Ejecutar T-A-S (k=1) en ≥200 con logging	JSONL sanitizado + Parquet por run; CoT solo en `logs_local/`	6	This	S2-01,02,04,09	M	Parquet `tas_200.parquet`; sin fugas de CoT en compartidos
S2-06	Ejecutar T-A-S+MAMV en ≥200	Igual a S2-05 pero con MAMV activo	8	This	S2-03,05	M	Parquet `mamv_200.parquet`; campos de votos presentes
S2-07	Embeddings locales (coherencia T→S)	Modelo `all-mpnet-base-v2`; cosine; sin CoT	5	José	S1 analytics	M	Cache local; función `coherence_ts()` con test
S2-08	Escribir coherencia a Parquet	Métricas agregadas (mean, p50, p90) y por ítem	4	José	S2-07	B	`coherence.parquet` legible y unido por `problem_id`
S2-09	Token caps por ítem y budget monitor por sprint	Límite ≤8k/ítem; reporte ≤1.5× baseline (generación)	4	This	S1 tokens	M	Alerta al acercarse al 90% del presupuesto; tabla de consumo
S2-10	McNemar y KPIs: base vs T-A-S; base vs T-A-S+MAMV	Tabla ΔAcc, p-values crudos, tokens (generación vs embeddings)	5	José	S2-05,06,08,09	M	`metrics_s2.parquet` + tabla en Sprint2.md con interpretación
S2-11	Taxonomía de errores (pipeline + primeras etiquetas)	Categorías: aritmética, interpretación, ruptura, formato	4	José	S2-05,06	M	50 ejemplos rotulados; conteos por categoría en Parquet
S2-12	Safety audit: sanitización extendida	Validar que nuevos campos (votos, coherencia) no exponen PII/CoT	6	Lorena	S2-05..08	M	Checklist firmado; tests de `sanitize()` pasan
S2-13	Unit tests MAMV/jitter	Cobertura básica de ramas y desempates	4	Julio	S2-03	B	`pytest -q` OK en CI
S2-14	Unit tests embeddings/coherencia	Casos deterministas y de borde	2	José	S2-07	B	`pytest -q` OK; tolerancias documentadas
S2-15	Unit tests sanitización campos nuevos	Falla si aparece texto sensible en JSONL compartido	2	Lorena	S2-12	B	Test detecta fuga simulada y bloquea write
S2-16	Actualizar README (flags/params MAMV, coherencia)	Sección de uso y reproducibilidad	4	Valeria	S2-03..10	B	Usuario externo reproduce corridas siguiendo README
S2-17	Sprint2.md (reporte)	Resultados, tablas, interpretación y próximos pasos	4	Valeria	S2-10,11	B	`/reports/Sprint2.md` con KPIs y nota metodológica
S2-18	Micro-tuning rápido (validación 50 ítems)	Smoke test con Tesis=0.6 vs 0.7; elección final documentada	4	Julio	S2-02	M	Tabla comparativa y decisión en Sprint2.md
S2-19	Calidad de datos & triage	Revisión de fallos/tiempos; lista de ítems “deferred”	3	José	S2-05,06	B	Registro de incidentes y re-ejecuciones controladas

S3

ID	Task	DoD	Est. (h)	Owner	Dep.	Riesgo	Aceptación
S3-01	Incorporar corpus de debate en repo con licencia Apache-2.0	`/prompts/debate/` con README, plantillas T, A, S versionadas	5	This	—	B	Carpeta creada, licencia incluida, ejemplos mínimos validados
S3-02	Lint de prompts de debate (deny-list, regex, longitud) y CI	Script `lint_prompts.py` y job en CI; falla si hay PII o CoT	4	Lorena	S3-01	M	CI rojo ante prompt con PII; reporte de lint en artefactos
S3-03	Loader TruthfulQA y normalización de respuestas	Funciones de carga y parsing; mapping yes/no y libre corta	7	José	—	M	Tests de parsing; muestra de 200 ítems con seed fija
S3-04	Verificación set GSM8K de 200 (el mismo de S1)	Hash de IDs coincide; tabla de referencia versionada	2	José	—	B	`ids_gsm8k_200.json` igual a S1; checksum registrado
S3-05	T-A-S (k=1) con debate en GSM8K 200	JSONL sanitizado y Parquet por run; sin fuga de CoT	6	This	S3-01,02,04	M	`tas_gsm8k_200.parquet` generado y legible
S3-06	T-A-S+MAMV (k=1) con debate en GSM8K 200	3 instancias con jitter Tesis {0.65, 0.70, 0.75}; mayoría simple	6	This	S3-05	M	`mamv_gsm8k_200.parquet` con votos por instancia
S3-07	T-A-S (k=1) con debate en TruthfulQA 200	Misma estructura de logging; guardarrail de formato	6	This	S3-01,02,03	M	`tas_tqa_200.parquet` disponible
S3-08	T-A-S+MAMV (k=1) en TruthfulQA 200	Igual que S3-06, aplicado a TQA	6	This	S3-07	M	`mamv_tqa_200.parquet` con trazas de votos
S3-09	Métricas de coherencia T→S con embeddings locales	`all-mpnet-base-v2`, cosine; sin CoT; parquet de coherencia	6	José	S3-05..08	M	`coherence_tas.parquet` por dataset y variante
S3-10	Runner k=2 en GSM8K subset 100 con regla de futilidad	Freno intermedio en 50; aborta si ΔAcc<+2pp o costo>2.5× o contradicción>baseline+2pp	8	Julio	S3-05	M	Log de decisión de futilidad; `k2_gsm8k_100.parquet`
S3-11	Monitor de costo y alertas	Cap por ítem ≤8k; alertas a 2.5× y 2.8×; embeddings separados	3	This	S3-05..08	B	Tabla de consumo con alertas y resumen por run
S3-12	McNemar y KPIs en GSM8K	Baseline vs T-A-S; baseline vs T-A-S+MAMV; p-values crudos	5	José	S3-05,06	M	`metrics_gsm8k_s3.parquet` y tabla en Sprint3.md
S3-13	McNemar y KPIs en TruthfulQA	Misma metodología que S3-12 para TQA	5	José	S3-07,08	M	`metrics_tqa_s3.parquet` y tabla en Sprint3.md
S3-14	Coherencia inter-iteración k=2	c(T1,S1), c(S1,S2), c(T1,S2), deltas y contradicción	4	José	S3-10	M	`coherence_k2.parquet` con estadísticas resumen
S3-15	Taxonomía de errores ampliada	Etiquetado mínimo: 50 GSM8K y 50 TQA adicionales	4	José	S3-05..08	M	Parquet de etiquetas y conteos por categoría
S3-16	Safety audit S3 (prompts, nuevos campos, Parquet)	Checklist firmado; pruebas de sanitización extendida	5	Lorena	S3-01..09	M	Tests bloquean cualquier fuga simulada; checklist en repo
S3-17	Tests unitarios S3	Lint de prompts, TQA parsing, futilidad k=2, coherencia k=2	5	Julio	S3-02,03,10,14	B	`pytest -q` pasa en CI
S3-18	CLI para S3	Comando con flags: dataset, k, mamv, debate, seed, budget	4	This	S3-05..08	B	`tas run --dataset tqa --k 1 --mamv on` funcional
S3-19	README actualizado	Uso de debate, TruthfulQA, k=2, criterios de adopción	4	Valeria	S3-01..18	B	Usuario externo reproduce corridas siguiendo README
S3-20	Sprint3.md (Informe de transferencia)	Resultados, tablas, interpretación, figuras simples	6	Valeria	S3-12..15	B	`/reports/Sprint3.md` con conclusiones claras

S4

ID	Task	DoD	Est. (h)	Owner	Dep.	Riesgo	Aceptación
S4-01	Triage de S3 y decisión k=2 (principal vs anexo)	Acta técnica con criterios: ≥+2 pp, coherencia OK, coste ≤3.0×	4	José	S3 resultados	M	Documento “k2_decision.md” firmado por SM
S4-02	Consolidar tablas finales (GSM8K + TQA)	Esquema unificado: baseline / T-A-S / MAMV [/ k=2 si aplica]	6	José	S3-12..13	M	`results/*.{parquet,csv}` con schema válido
S4-03	McNemar y KPIs finales por dataset y variante	ΔAcc, p-values crudos, guardarrail de formato; α=0.05	8	José	S4-02	M	`results/kpi_final.parquet` y tabla en Sprint4.md
S4-04	Ablations: MAMV ON/OFF y Debate ON/OFF	2×2 por dataset; costo/tokens por celda	6	Julio	S4-02	M	Tabla “ablation_2x2” por dataset con lectura clara
S4-05	No-regresión (invalid/format)	Comparado a baseline: ≤ +2 pp	3	José	S4-02	B	Tabla de control “format_guardrail.csv”
S4-06	Figuras ΔAcc vs costo (barras pareadas)	Estilo formal con acento morado; export PNG a `/releases/v1.0/figs/`	6	José	S4-03	M	`fig_acc_cost_{dataset}.png` verificados
S4-07	Curvas de consumo y caps (tokens)	Incluye umbrales y alertas	4	José	S4-03	B	`fig_tokens_{dataset}.png`
S4-08	Heatmaps coherencia/contradicción	Umbral contradicción y medias/deltas	5	José	S4-02	M	`fig_coherence_{dataset}.png`
S4-09	Distribución de categorías de error + ejemplos	Top-k categorías y 1–2 ejemplos por categoría	5	José	S4-02	M	`errors_{dataset}.csv` + sección en Sprint4.md
S4-10	Safety audit final (sanitización, CoT, prompts)	Checklist firmado; CI con lint pasa	8	Lorena	S4-02..09	M	`safety_checklist_s4.md` + CI verde

S4-11	SECURITY.md (placeholder contacto)	Plantilla con proceso de reporte	2	Lorena	—	B	`SECURITY.md` en raíz
S4-12	CODE_OF_CONDUCT.md	Plantilla adoptada	2	Lorena	—	B	Archivo en raíz
S4-13	Data Card & Model Card	Alcance, datos, métricas, límites, uso responsable	6	Valeria (+Lorena rev)	S4-03..09	M	`data_card.md` + `model_card.md`
S4-14	Replication pack (`run_all.sh`, `README_repl.md`)	Reproduce resultados principales sin CoT (seeds fijos)	8	This	S4-02..09	M	Script ejecuta fin-a-fin en limpio
S4-15	Dry-run de replicación (headless)	Log de ejecución + tiempos; sin errores	6	This	S4-14	M	`replication_log.txt` sin fallos
S4-16	Empaquetado `/releases/v1.0/`	Árbol completo, nombres acordados	6	This	S4-06..09,14	M	Estructura final creada y auditada
S4-17	CITATION.cff (DOI placeholder)	Formato válido con autores/afiliaciones placeholders	2	Valeria	—	B	`CITATION.cff` validado
S4-18	Paper de replicación (Draft)	Secciones: método, exp., resultados, ablations, ética/limitaciones	16	Valeria	S4-03..09	M	`/releases/v1.0/paper/draft.pdf`
S4-19	Exportar figuras para paper	Resolución lista para impresión	3	José	S4-06..08	B	PNG/SVG en `paper/figs/`
S4-20	README final (raíz)	Instrucciones, licencias, reproducibilidad	4	Valeria	S4-14..16,18	B	PR aprobado; lectura “fresh-install”
S4-21	Release notes y CHANGELOG	Cambios clave y compatibilidad	3	This	S4-16	B	`CHANGELOG.md` + notas de release
S4-22	Preparar metadata Zenodo (placeholders)	Título, autores (placeholder), descripción, licencias	4	This	S4-16,17,20	M	JSON/GUI completado; lista para publicar
S4-23	QA final (checklist integral)	Revisión cruzada (SM + Safety)	4	This + Lorena	S4-10..22	M	`qa_final.md` sin pendientes críticos
S4-24	Executive summary (1 página)	Resumen para managers (formal, acento morado)	4	Valeria	S4-03..09	B	`paper/executive_summary.pdf`
S4-25	Preprint skeleton (no envío)	Carpeta con fuentes y guía de envío	5	Valeria	S4-18	B	`/paper/preprint/` preparado