Sprints

Ricardo Horta - Julio de Aquino - José Pech - Valeria Hernández - Lorena Pérez

---

Sprint 1 - Plan (20–26 oct 2025)

Horario de trabajo: lun–vie 10:00–18:00.
Sprint Goal - Infrastructure First:
Infra reproducible con `uv`, flujo Prefect **Tesis→Antítesis→Síntesis (k=1)**, esquema de **logging JSONL** + **Parquet** y **harness de evaluación** (incl. **McNemar**). **Baseline** completo en **≥200** problemas GSM8K; **T-A-S** ejecutado en **piloto (~50)** para validar fin-a-fin.

Capacidad (supuesto 80% efectiva ≈ 32 h/persona)

Persona	Rol	Cap. efectiva (h)	Horas planificadas	Carga
This au Chocolat	Scrum Master + Orchestration	32	31	97%
Julio de Aquino	MLE	32	21	66%
José Pech	Data / Evaluación	32	25	78%
Lorena Pérez	AI Safety & Compliance	32	12	38%
Valeria Hernández	Tech Writing	32	10	31%
Total asignado: 99 h (holgura suficiente para contingencias y revisiones).				

Sprint Backlog (S1)

Convenciones:

- **DoD** = Definition of Done; **Riesgo:** B/M/A (Bajo/Medio/Alto).
- **Aceptación** = criterios verificables.

ID	Task	DoD	Estimate (h)	Ow
S1-01	Inicializar repo con `uv` y estructura de carpetas	`pyproject.toml`, `uv.lock`, árbol base creado; `.gitignore` para `logs_local/`	6	Thi
S1-02	Configuración CI (lint/tests) + pre-commit	Workflow GH Actions (lint+tests) pasa; hooks activos	2	Thi
S1-03	Loader GSM8K + normalización	Función `load_batch(n, seed)` y normalización de respuestas numéricas	6	Jos
S1-04	Evaluador exact-match (GSM8K)	`evaluate_exact_match(y_true, y_pred)` con normalización	4	Jos
S1-05	Contador de tokens (prompt+completion)	`count_tokens(event)` integrado al logger	3	Thi
S1-06	Baseline runner ≥200 (single-call)	Flow Prefect que ejecuta baseline con logging JSONL	8	Thi
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	Juli
S1-08	Prefect flow T-A-S (orquestación)	Tareas encadenadas T→A→S (k=1) con manejo de errores	6	Thi
S1-09	Logger JSONL + sanitización/anonimización	`log_event_jsonl(record)` usa `sanitize()` y hashing con sal	6	Lor
S1-10	Agregación Parquet (analytics)	Job que convierte JSONL→Parquet por run	4	Jos
S1-11	Unit tests mínimos (parser/evaluator/sanitize/flow)	≥4 tests críticos; cobertura básica	8	Juli (4) Jos (3) Lor (1)
S1-12	Pilot run T-A-S (~50)	Ejecución real; logs locales y eventos compartidos generados	4	Juli
S1-13	McNemar + KPIs (baseline vs T-A-S)	Script estadístico; tabla de métricas; cálculo ΔAcc y tokens	5	Jos
S1-14	README (cómo correr y reproducir)	Secciones: setup `uv`, baseline, T-A-S, métricas	4	Val

ID	Task	DoD	Estimate (h)	Ow
S1-15	Reporte corto Sprint1.md	Resumen de resultados, riesgos, próximos pasos	3	Val
S1-16	Templating de prompts + hashing de prompt/resp	Funciones utilitarias y pruebas	3	Juli

Nice-to-have (no comprometido con el Goal)

ID	Task	Estimate (h)	Owner	Nota
N1	MAMV prototipo (3 instancias, votación mayoría)	6	Julio	Integrable como “estrategia de síntesis alternativa”
N2	CLI mínima `tas run --dataset gsm8k --n 50`	3	This	Azúcar sintáctico sobre Prefect flow
N3	Métrica de “coherencia T→S” (similitud textual)	4	José	Para S2, si da tiempo

---

Pequeño reporte: qué se hará en el Sprint 1

Alcance:

1. Montar la columna vertebral del proyecto: entorno reproducible con `uv`, estructura de repo, CI, y orquestación Prefect.
2. Implementar el **baseline** de referencia (single-call) y correrlo en **≥200** problemas GSM8K con logging consistente (JSONL) y curado a **Parquet** para análisis.
3. Implementar el **pipeline mínimo T-A-S (k=1)** con control de temperatura (0.7/0.2), resguardando el **Chain-of-Thought solo en local** y sanitizando todo evento compartido.
4. Ejecutar un **piloto T-A-S (~50 ítems)** para validar end-to-end y confirmar que las métricas, conteo de tokens y logs se generan correctamente.
5. Calcular **KPIs** (ΔAcc vs. tokens) y **McNemar** sobre el subset medido, documentar resultados y dejar **README** y **Sprint1.md** listos en el repositorio.

Entregables al cierre del sprint:

- Repo con `uv.lock`, estructura, CI y tests mínimos.
- Artefactos de ejecución: JSONL (eventos) y Parquet (analítica).

- Baseline (≥200) y piloto T-A-S (~50) ejecutados y registrados.
- Script/nota técnica de **McNemar** y KPIs (tabla y breve interpretación).
- **README** operativo y **/reports/Sprint1.md** con resultados y próximos pasos.

Riesgos principales y mitigación:

- **Costo en tokens** (M): límite por ítem y sampleo inicial de 50 en T-A-S.
- **Desviación de formato/parseo** (M): normalización estricta de respuestas GSM8K + tests.
- **Filtración de CoT** (M): separación `logs_local/` (gitignored) vs `logs/events/` y `SAVE_COT_LOCAL_ONLY=true` obligatorio.
- **Carga desigual** (B): backlog conservador y holgura para redistribuir horas si hay bloqueos.

Artefactos y ubicación:

- Código y issues: **GitHub repo**.
- Resultados: `/logs/events/*.jsonl`, `/analytics/parquet/*.parquet`.
- Documentación: `/README.md`, `/reports/Sprint1.md`.

Sprint 2 - Plan (27 oct–2 nov 2025)

Horario: lun–vie 10:00–18:00
Sprint Goal: alcanzar **Éxito mínimo** (ΔAcc ≥ **+5 pp** con **≤2.5×** tokens) en **≥200** ítems **GSM8K**, con **McNemar p < 0.05**, comparando **baseline vs. T-A-S (k=1)** y **baseline vs. T-A-S+MAMV (3 instancias)**; además, **análisis cualitativo** (taxonomía de errores) y **coherencia T→S por embeddings locales**.

Capacidad (80% efectiva ≈ 32 h/persona)

Persona	Rol	Cap. efectiva (h)	Horas planificadas	Carga
This au Chocolat	Scrum Master + Orchestration	32	28	88%
Julio de Aquino	MLE	32	21	66%
José Pech	Data / Evaluación	32	31	97%
Lorena Pérez	AI Safety & Compliance	32	11	34%

Persona	Rol	Cap. efectiva (h)	Horas planificadas	Carga
Valeria Hernández	Tech Writing	32	10	31%
Total asignado: 101 h (holgura para bloqueos y re-intentos).				

Sprint Backlog (S2)

Convenciones: DoD = Definition of Done; Riesgo: B/M/A; Aceptación = criterios verificables.

ID	Task	DoD	Est. (h)	Owner	Dependencias	Riesg
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
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
S2-03	Implementar MAMV (3 instancias)	Síntesis por mayoría simple con trazabilidad por instancia	10	Julio	S2-02	M
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
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
S2-06	Ejecutar T-A-S+MAMV en ≥200	Igual a S2-05 pero con MAMV activo	8	This	S2-03,05	M

ID	Task	DoD	Est. (h)	Owner	Dependencias	Riesg
S2-07	Embeddings locales (coherencia T→S)	Modelo `all-mpnet-base-v2` ; cosine; sin CoT	5	José	S1 analytics	M
S2-08	Escribir coherencia a Parquet	Métricas agregadas (mean, p50, p90) y por ítem	4	José	S2-07	B
S2-09	Token caps por ítem y budget monitor por sprint	Límite ≤8k/ítem; reporte ≤1.5× baseline (generación)	4	This	S1 tokens	M
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
S2-11	Taxonomía de errores (pipeline + primeras etiquetas)	Categorías: aritmética, interpretación, ruptura, formato	4	José	S2-05,06	M
S2-12	Safety audit: sanitización extendida	Validar que nuevos campos (votos, coherencia) no exponen PII/CoT	6	Lorena	S2-05..08	M
S2-13	Unit tests MAMV/jitter	Cobertura básica de ramas y desempates	4	Julio	S2-03	B
S2-14	Unit tests embeddings/coherencia	Casos deterministas y de borde	2	José	S2-07	B
S2-15	Unit tests sanitización campos nuevos	Falla si aparece texto sensible en JSONL compartido	2	Lorena	S2-12	B
S2-16	Actualizar README (flags/params MAMV,	Sección de uso y	4	Valeria	S2-03..10	B

ID	Task	DoD	Est. (h)	Owner	Dependencias	Riesg
	coherencia)	reproducibilidad				
S2-17	Sprint2.md (reporte)	Resultados, tablas, interpretación y próximos pasos	4	Valeria	S2-10,11	B
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
S2-19	Calidad de datos & triage	Revisión de fallos/tiempos; lista de ítems “deferred”	3	José	S2-05,06	B

Nice-to-have (no comprometido con el Goal)

ID	Task	Est. (h)	Owner	Nota
S2-N1	**CLI** `tas run …` (azúcar sobre Prefect)	3	This	Solo si hay tiempo
S2-N2	Mini-notebook de visualización (histogramas/boxplots)	3	José	Complemento para Sprint2.md

---

Reporte breve: ¿Qué haremos y cómo mediremos?

1) Escalado y comparaciones emparejadas
Ejecutaremos **T-A-S (k=1)** y **T-A-S+MAMV (3 instancias)** sobre **los mismos 200** problemas usados en S1 (seed=42, estratos). Así garantizamos **emparejamiento 1-a-1** para las pruebas estadísticas.

2) Control de costos y robustez operativa
El runner de Prefect tendrá **reintentos con backoff** y control de **rate-limit**. Imponemos **límite ≤8k tokens/ítem** y un **budget ≤1.5×** del baseline para **generación**; los **embeddings** se reportan **por separado** (no entran al cap de generación).

3) Métricas y estadística
Calcularemos **ΔAcc** y **tokens** (prompt, completion, total de generación), además de **coherencia T→S** con **embeddings locales** (`all-mpnet-base-v2`, cosine). Aplicaremos **McNemar** para (a) **baseline vs T-A-S** y (b) **baseline vs T-A-S+MAMV** y reportaremos **p-values crudos** (sin corrección múltiple, con **nota metodológica** sobre α=0.05). Si el ΔAcc alcanza **+5 pp** con **≤2.5×** tokens y **p<0.05**, cumplimos **Éxito mínimo**.

4) Safety y trazabilidad
El **Chain-of-Thought** permanecerá **solo en local** (`logs_local/`, gitignored). En JSONL/Parquet compartidos únicamente quedarán **resúmenes**, **métricas**, **hashes** y campos sanitizados. Lorena validará sanitización extendida para nuevos campos (votos MAMV, coherencia).

5) Análisis cualitativo
Construiremos una **taxonomía de errores** (aritmética, interpretación, ruptura de razonamiento, formato) con una primera muestra **(≥50)** para entender dónde aporta más la dialéctica/MAMV. Estos conteos y ejemplos guiados se incluirán en **Sprint2.md**.

6) Documentación
Actualizaremos el **README** (parámetros, reproducibilidad) y elaboraremos **Sprint2.md** con KPIs, p-values, consumo de tokens (separado por generación vs embeddings), análisis de errores y decisiones de micro-tuning.

Sprint 3 - Generalización + Debate

Fechas: 3–9 nov 2025
Horario: 10:00–18:00
Objetivo: integrar el corpus de debate en el flujo de inferencia y demostrar transferencia a TruthfulQA, manteniendo GSM8K como base. Se comparan formalmente baseline vs T-A-S (k=1) y baseline vs T-A-S+MAMV (k=1) en GSM8K y TruthfulQA. Se ejecuta k=2 en un subset exploratorio con regla de futilidad.

Capacidad (80% efectiva ≈ 32 h por persona)

Persona	Rol	Cap. efectiva (h)	Horas planificadas	Carga
This au Chocolat	Scrum Master + Orchestration	32	29	91%
Julio de Aquino	MLE	32	25	78%

Persona	Rol	Cap. efectiva (h)	Horas planificadas	Carga
José Pech	Data / Evaluación	32	31	97%
Lorena Pérez	AI Safety & Compliance	32	11	34%
Valeria Hernández	Tech Writing	32	10	31%
Total asignado: 106 h				

Sprint Backlog (S3)

Convenciones: DoD = Definition of Done; Riesgo: B/M/A; Aceptación = criterios verificables.

ID	Task	DoD	Est. (h)	Owner	Dep.	Riesgo
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
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
S3-03	Loader TruthfulQA y normalización de respuestas	Funciones de carga y parsing; mapping yes/no y libre corta	7	José	—	M
S3-04	Verificación set GSM8K de 200 (el mismo de S1)	Hash de IDs coincide; tabla de referencia versionada	2	José	—	B
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

ID	Task	DoD	Est. (h)	Owner	Dep.	Riesgo
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
S3-07	T-A-S (k=1) con debate en TruthfulQA 200	Misma estructura de logging; guardarraíl de formato	6	This	S3-01,02,03	M
S3-08	T-A-S+MAMV (k=1) en TruthfulQA 200	Igual que S3-06, aplicado a TQA	6	This	S3-07	M
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
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
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
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
S3-13	McNemar y KPIs en TruthfulQA	Misma metodología que S3-12 para TQA	5	José	S3-07,08	M
S3-14	Coherencia inter-iteración k=2	c(T1,S1), c(S1,S2), c(T1,S2), deltas y contradicción	4	José	S3-10	M
S3-15	Taxonomía de errores ampliada	Etiquetado mínimo: 50 GSM8K y 50 TQA adicionales	4	José	S3-05..08	M
S3-16	Safety audit S3 (prompts, nuevos	Checklist firmado; pruebas de sanitización extendida	5	Lorena	S3-01..09	M

ID	Task	DoD	Est. (h)	Owner	Dep.	Riesgo
	campos, Parquet)					
S3-17	Tests unitarios S3	Lint de prompts, TQA parsing, futilidad k=2, coherencia k=2	5	Julio	S3-02,03,10,14	B
S3-18	CLI para S3	Comando con flags: dataset, k, mamv, debate, seed, budget	4	This	S3-05..08	B
S3-19	README actualizado	Uso de debate, TruthfulQA, k=2, criterios de adopción	4	Valeria	S3-01..18	B
S3-20	Sprint3.md (Informe de transferencia)	Resultados, tablas, interpretación, figuras simples	6	Valeria	S3-12..15	B

Nice-to-have (no comprometido con el objetivo)

ID	Task	Est. (h)	Owner	Nota
S3-N1	Notebook de visualización de transferencia	3	José	Histogramas y boxplots de coherencia y ΔAcc
S3-N2	Loader ARC (esqueleto)	4	José	Dejar listo para S4 si se decide ampliar

---

Reporte breve: qué se hará y cómo se medirá

1. **Debate integrado al flujo de inferencia**
   Se añaden plantillas de prompts para Tesis, Antítesis y Síntesis. El lint automático en CI evita PII, patrones prohibidos y errores de formato.
2. **Generalización a TruthfulQA**
   Se ejecutan T-A-S y T-A-S+MAMV con k=1 en 200 ítems de TruthfulQA, replicando la metodología de GSM8K. Se comparan formalmente contra el baseline con McNemar (p-values crudos, α=0.05).
3. **k=2 exploratorio con regla de futilidad**
   En GSM8K subset 100 se prueba k=2. En la mitad de la corrida (50 ítems) se detiene si

hay baja ganancia de exactitud, costo excesivo o contradicción elevada. La meta es no gastar más del 50% del presupuesto si la ganancia es mínima.

4. **Métricas y umbrales**
   Exactitud, costo en tokens de generación con caps por ítem y por corrida, embeddings reportados por separado, coherencia T→S con embeddings locales y coherencia inter-iteración para k=2. Se mantiene el guardarraíl de formato.
5. **Safety y reproducibilidad**
   CoT únicamente en local, logs sanitizados, licencias claras, scripts de lint y tests unitarios. CLI para reproducir corridas con flags.
6. **Entregables**
   Parquets de resultados para GSM8K y TQA; tablas de McNemar y KPIs; parquet de coherencia y de k=2; taxonomía de errores ampliada; README actualizado; Informe de transferencia en Sprint3.md.

¡Perfecto! Con tu autorización para **S4 (Auditoría y Hardening)**, aquí está el plan completo — capacidad, backlog y reporte breve— alineado a tus reglas de éxito y publicación v1.0.

Sprint 4 — Auditoría y Hardening

Fechas: 10–16 nov 2025
Horario: 10:00–18:00
Sprint Goal: auditoría estadística final, reproducibilidad end-to-end, hardening de AI Safety, y **paquete público v1.0** (GitHub Release + DOI Zenodo) con ablations claras (**MAMV ON/OFF** y **Debate ON/OFF**) y costos/tokens; k=2 en tabla principal **solo** si S3 mostró adopción viable, si no va a anexo.

Capacidad (80% efectiva ≈ 32 h/persona)

Persona	Rol	Cap. efectiva (h)	Horas planificadas	Carga
This au Chocolat	Scrum Master + Orchestration	32	33	103%
Julio de Aquino	MLE	32	14	44%
José Pech	Data / Evaluación	32	46	144%
Lorena Pérez	AI Safety & Compliance	32	12	38%
Valeria Hernández	Tech Writing	32	23	72%

> Nota: José concentra gran parte del análisis final. Si se observa presión, podemos mover 6–8 h de generación de figuras a Lorena o Julio. Total planificado ≈ **128 h** (dentro de la capacidad global del equipo, redistribuible si hay bloqueos).

Sprint Backlog (S4)

Convenciones: **DoD** = Definition of Done; **Riesgo**: B/M/A; **Aceptación** = criterios verificables.

ID	Task	DoD	Est. (h)	Owner	Dep.
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 resultado**
S4-02	**Consolidar tablas finales (GSM8K + TQA)**	**Esquema unificado: baseline / T-A-S / MAMV [/ k=2 si aplica]**	**6**	**José**	**S3-12..13**
S4-03	**McNemar y KPIs finales por dataset y variante**	**ΔAcc, p-values crudos, guardarraíl de formato; α=0.05**	**8**	**José**	**S4-02**
S4-04	**Ablations: MAMV ON/OFF y Debate ON/OFF**	**2×2 por dataset; costo/tokens por celda**	**6**	**Julio**	**S4-02**
S4-05	**No-regresión (invalid/format)**	**Comparado a baseline: ≤ +2 pp**	**3**	**José**	**S4-02**
S4-06	**Figuras ΔAcc vs costo (barras pareadas)**	**Estilo formal con acento morado; export PNG a `/releases/v1.0/figs/`**	**6**	**José**	**S4-03**
S4-07	**Curvas de consumo y caps (tokens)**	**Incluye umbrales y alertas**	**4**	**José**	**S4-03**
S4-08	**Heatmaps coherencia/contradicción**	**Umbral contradicción y medias/deltas**	**5**	**José**	**S4-02**
S4-09	**Distribución de categorías de error + ejemplos**	**Top-k categorías y 1–2 ejemplos por categoría**	**5**	**José**	**S4-02**
S4-10	**Safety audit final (sanitización, CoT, prompts)**	**Checklist firmado; CI con lint pasa**	**8**	**Lorena**	**S4-02..09**

ID	Task	DoD	Est. (h)	Owner	Dep.
S4-11	SECURITY.md (placeholder contacto)	Plantilla con proceso de reporte	2	Lorena	—
S4-12	CODE_OF_CONDUCT.md	Plantilla adoptada	2	Lorena	—
S4-13	Data Card & Model Card	Alcance, datos, métricas, límites, uso responsable	6	Valeria (+Lorena rev)	S4-03..09
S4-14	Replication pack (`run_all.sh`, `README_repl.md`)	Reproduce resultados principales sin CoT (seeds fijos)	8	This	S4-02..09
S4-15	Dry-run de replicación (headless)	Log de ejecución + tiempos; sin errores	6	This	S4-14
S4-16	Empaquetado `/releases/v1.0/`	Árbol completo, nombres acordados	6	This	S4-06..09,14
S4-17	CITATION.cff (DOI placeholder)	Formato válido con autores/afiliaciones placeholders	2	Valeria	—
S4-18	Paper de replicación (Draft)	Secciones: método, exp., resultados, ablations, ética/limitaciones	16	Valeria	S4-03..09
S4-19	Exportar figuras para paper	Resolución lista para impresión	3	José	S4-06..08
S4-20	README final (raíz)	Instrucciones, licencias, reproducibilidad	4	Valeria	S4-14..16,18
S4-21	Release notes y CHANGELOG	Cambios clave y compatibilidad	3	This	S4-16
S4-22	Preparar metadata Zenodo (placeholders)	Título, autores (placeholder), descripción, licencias	4	This	S4-16,17,20
S4-23	QA final (checklist integral)	Revisión cruzada (SM + Safety)	4	This + Lorena	S4-10..22
S4-24	Executive summary (1 página)	Resumen para managers (formal, acento morado)	4	Valeria	S4-03..09
S4-25	Preprint skeleton (no envío)	Carpeta con fuentes y guía de envío	5	Valeria	S4-18

Nomenclatura acordada (resultados):
`results/{gsm8k|tqa}__{baseline|tas|mamv|k2}.parquet`
`figs/fig_{id}_{dataset}_{variante}.png`
Estilo: formal con acento **morado**.

Reporte breve: qué se cierra y qué se publica

1) Validación final contra el criterio global

- Verificamos **Éxito mínimo** (ΔAcc ≥ +5 pp con ≤2.5× tokens de **generación**) en **≥1** dataset y **No-regresión** en el otro (ΔAcc ≥ 0 pp, invalid/format ≤ baseline + 2 pp).
- Etiquetamos **Target+** si **ambos** alcanzan Éxito mínimo.

2) Resultados y ablations con costos

- Publicamos tablas y figuras **ΔAcc vs costo** por dataset/variante y **2×2** de ablations (**MAMV ON/OFF**, **Debate ON/OFF**).
- Si S3 avaló **k=2**, aparece en tabla principal; si no, queda en **anexo**.

3) Reproducibilidad

- Entregamos `replication_pack/run_all.sh` y `README_repl.md` para recrear resultados principales con **seeds fijos** y **sin CoT**; hacemos un **dry-run** completo y registramos logs/tiempos.

4) AI Safety & Compliance

- **CoT** no se publica; JSONL/Parquet sanitizados; prompts con lint en CI; **SECURITY.md**, **CODE_OF_CONDUCT.md** y **checklist de safety** firmados.

5) Publicación v1.0

- Estructura `/releases/v1.0/` con **resultados finales**, **figuras**, **draft del paper**, **licencias** y **CITATION.cff** con **DOI placeholder** (Zenodo).
- **Preprint**: se deja **skeleton** listo (no envío en S4).

6) Documentación y comunicación

- **README** final en raíz, **Release notes/CHANGELOG**, **Executive summary** de 1 página y **Data/Model Card** (alcance, datos, métricas, límites y uso responsable).