SprintsV3

SprintsV3 — Plan Reducido para Entrega

Contexto: Nos estamos quedando sin tiempo. Este documento **descarta todo lo innecesario** y se concentra únicamente en **cerrar lo crítico** para entregar el proyecto con resultados sólidos.

Principio rector: Si no es indispensable para demostrar que el método funciona (o no) y documentarlo formalmente, **no lo hacemos**.

Estado actual (fin de S2)

✅ **Sprint 2 completado:**

- Baseline: **98% accuracy**, 15.8k tokens
- T-A-S (k=1): **96% accuracy**, 255k tokens (-2pp, 16× más caro)
- T-A-S+MAMV: **98% accuracy**, 758k tokens (iguala baseline pero **47× más caro**)

Conclusión S2: MAMV no es costo-efectivo. El método funciona técnicamente pero no mejora sobre baseline en GSM8K.

Deuda técnica identificada:

- `gsm8k_0029` falla consistentemente en todos los modelos
- Una corrida MAMV tuvo 44/50 fallos por problemas de conexión (documentado pero no crítico)
- Falta análisis estadístico formal (McNemar, p-values)
- Falta transferencia a otro dataset

Sprint 3 LEAN — Validación Formal + Transferencia

Fechas: 3–9 nov 2025
Objetivo: Demostrar transferencia a TruthfulQA y cerrar análisis estadístico formal. **Eliminar k=2** (ya sabemos que no vale la pena por costo).

Eliminaciones vs Sprints.md original

❌ **ELIMINADO:** S3-01 (corpus de debate) — no tenemos tiempo de validar debate
❌ **ELIMINADO:** S3-02 (lint de prompts) — nice-to-have, no crítico
❌ **ELIMINADO:** S3-09 (coherencia T→S con embeddings) — métrica exploratoria
❌ **ELIMINADO:** S3-10 (k=2 con futilidad) — demasiado caro, S2 ya mostró el problema
❌ **ELIMINADO:** S3-11 (monitor de costo) — ya tenemos métricas en notebooks
❌ **ELIMINADO:** S3-14 (coherencia inter-iteración k=2) — depende de k=2 eliminado
❌ **ELIMINADO:** S3-16 (safety audit S3) — consolidamos en S4
❌ **ELIMINADO:** S3-18 (CLI) — funcionalidad existe, no es bloqueante
❌ **ELIMINADO:** S3-N1 y S3-N2 (nice-to-have)

Backlog S3 REDUCIDO

ID	Task	DoD	Est. (h)	Owner	Dep.	Riesgo
S3-03	Loader TruthfulQA y normalización	Funciones de carga; 200 ítems con seed fija	6	José	—	M
S3-04	Verificación set GSM8K 200	Hash de IDs coincide con S1/S2	2	José	—	B
S3-05	T-A-S (k=1) en GSM8K 200 (sin debate)	Parquet sanitizado, mismo método S2	5	This	S3-04	M
S3-06	T-A-S+MAMV (k=1) en GSM8K 200	3 instancias con jitter, mayoría simple	5	This	S3-05	M
S3-07	T-A-S (k=1) en TruthfulQA 200 (sin debate)	Misma estructura S3-05	5	This	S3-03	M
S3-08	T-A-S+MAMV (k=1) en TruthfulQA 200	Igual S3-06, dataset TQA	5	This	S3-07	M
S3-12	McNemar y KPIs GSM8K	Baseline vs T-A-S; baseline vs MAMV; p-values	5	José	S3-05,06	M
S3-13	McNemar y KPIs TruthfulQA	Misma metodología S3-12 para TQA	5	José	S3-07,08	M
S3-15	Taxonomía de errores GSM8K y TQA	50 ejemplos por dataset etiquetados	4	José	S3-05..08	M
S3-17	Tests unitarios TQA parsing	`pytest -q` pasa	3	Julio	S3-03	B
S3-19	README actualizado (TruthfulQA + k=1	Usuario externo reproduce corridas	3	Val	S3-03..15	B

ID	Task	DoD	Est. (h)	Owner	Dep.	Riesgo
	only)					
S3-20	Sprint3.md — Informe de transferencia	Resultados, tablas, interpretación clara	5	Val	S3-12..15	B

Total: ~53h planificadas (vs 106h en plan original)

Sprint 4 LEAN — Cierre y Publicación

Fechas: 10–16 nov 2025
Objetivo: Auditoría final, reproducibilidad y paquete público v1.0 **sin lujos**.

Eliminaciones vs Sprints.md original

❌ **ELIMINADO:** S4-01 (decisión k=2) — no hacemos k=2
❌ **ELIMINADO:** S4-04 (ablations debate ON/OFF) — no implementamos debate
❌ **ELIMINADO:** S4-07 (curvas de consumo) — figura nice-to-have
❌ **ELIMINADO:** S4-08 (heatmaps coherencia) — sin embeddings, no aplica
❌ **ELIMINADO:** S4-11 (SECURITY.md) — placeholder innecesario para v1.0 académico
❌ **ELIMINADO:** S4-12 (CODE_OF_CONDUCT.md) — innecesario para entrega académica
❌ **ELIMINADO:** S4-19 (exportar figuras) — se genera directo en paper
❌ **ELIMINADO:** S4-24 (executive summary) — no hay stakeholders externos
❌ **ELIMINADO:** S4-25 (preprint skeleton) — fuera de alcance temporal

Backlog S4 REDUCIDO

ID	Task	DoD	Est. (h)	Owner	Dep.	Riesgo
S4-02	Consolidar tablas finales (GSM8K + TQA)	Schema unificado: baseline / T-A-S / MAMV	5	José	S3-12,13	M
S4-03	McNemar y KPIs finales (ambos datasets)	ΔAcc, p-values, guardarraíl de formato	6	José	S4-02	M
S4-05	No-regresión (invalid/format)	≤ baseline + 2pp	2	José	S4-02	B

ID	Task	DoD	Est. (h)	Owner	Dep.	Riesgo
S4-06	Figuras ΔAcc vs costo (barras pareadas)	PNG formal para paper	5	José	S4-03	M
S4-09	Distribución categorías de error + ejemplos	Top-k con 1-2 ejemplos	4	José	S4-02	M
S4-10	Safety audit final (sanitización, CoT)	Checklist firmado; sin CoT en outputs	6	Lorena	S4-02..09	M
S4-13	Data Card & Model Card	Alcance, datos, métricas, límites	5	Val+Lorena	S4-03..09	M
S4-14	Replication pack (`run_all.sh`)	Reproduce principales sin CoT	6	This	S4-02..09	M
S4-15	Dry-run de replicación	Log sin errores	4	This	S4-14	M
S4-16	Empaquetado `/releases/v1.0/`	Estructura final auditada	4	This	S4-06..14	M
S4-17	CITATION.cff (DOI placeholder)	Formato válido	1	Val	—	B
S4-18	Paper de replicación (Draft)	Método, experimentos, resultados, limitaciones, ética	12	Val	S4-03..09	M
S4-20	README final (raíz)	Instrucciones, licencias, reproducibilidad	3	Val	S4-14..16,18	B
S4-21	Release notes y CHANGELOG	Cambios clave	2	This	S4-16	B
S4-22	Metadata Zenodo (placeholders)	Título, autores, descripción, licencias	3	This	S4-16,17,20	M
S4-23	QA final (checklist integral)	Revisión cruzada SM + Safety	3	This+Lorena	S4-10..22	M

Total: ~71h planificadas (vs 128h en plan original)

Criterios de Éxito para Entrega

Éxito Mínimo (necesario para aprobar)

En al menos 1 dataset:

- ΔAcc ≥ +5 pp sobre baseline
- Costo ≤ 2.5× tokens de **generación** vs baseline

En el otro dataset:

- ΔAcc ≥ 0 pp (no-regresión)
- Invalid/format ≤ baseline + 2pp

Target+ (excelencia)

Ambos datasets cumplen criterio de Éxito mínimo.

Realidad post-S2

- GSM8K: T-A-S **no cumple** (+5pp), tiene **-2pp** y es 16× más caro
- MAMV iguala baseline pero es **47× más caro** (fuera de budget)

Conclusión: Necesitamos que TruthfulQA muestre mejora, o el proyecto **no alcanza criterio de éxito**. S3 es **crítico**.

Capacidad Ajustada (solo S3+S4 reducidos)

Persona	S3 planificado (h)	S4 planificado (h)	Total (h)
This	20	22	42
José	27	32	59
Julio	3	0	3
Lorena	0	11	11
Valeria	8	23	31

Total: ~146h (vs 234h plan original completo)
Ahorro: 88h (~37% reducción)

Notas de Diseño Críticas

1. **No hacemos debate:** No hay tiempo de validar prompts dialécticos. Usamos mismo flujo T-A-S de S2.
2. **No hacemos k=2:** S2 demostró que MAMV con k=1 ya es 47× más caro sin beneficio. k=2 sería inviable.
3. **No hacemos coherencia con embeddings:** Métrica exploratoria, no crítica para demostrar efectividad del método.
4. **No hacemos CLI fancy:** La funcionalidad existe en scripts, suficiente para replicación.
5. **Enfoque en TruthfulQA:** Es nuestra última oportunidad de mostrar que el método funciona en al menos 1 dataset.
6. **Paper minimalista:** Draft con lo esencial. Sin executive summary, sin preprint skeleton. Solo resultados, método y discusión.
7. **Safety sin burocracia:** Sanitización de CoT (crítico) y checklist final. Sin SECURITY.md ni CODE_OF_CONDUCT.md innecesarios.

Entregables Finales v1.0

```
/releases/v1.0/
├── results/
│   ├── gsm8k__baseline.parquet
│   ├── gsm8k__tas.parquet
│   ├── gsm8k__mamv.parquet
│   ├── tqa__baseline.parquet
│   ├── tqa__tas.parquet
│   ├── tqa__mamv.parquet
│   └── kpi_final.parquet
├── figs/
│   ├── fig_acc_cost_gsm8k.png
│   ├── fig_acc_cost_tqa.png
│   └── fig_errors_distribution.png
├── paper/
│   └── draft.pdf
├── data_card.md
├── model_card.md
└── replication_pack/
    ├── run_all.sh
    └── README_repl.md
```

Raíz del repo:

- `README.md` (actualizado con TQA y reproducibilidad)
- `CITATION.cff` (DOI Zenodo placeholder)
- `CHANGELOG.md`
- `/reports/Sprint3.md` (informe de transferencia)

Riesgos Principales

Riesgo	Probabilidad	Impacto	Mitigación
TruthfulQA también muestra ΔAcc negativo	M	A	Documentar honestamente; ajustar narrativa
Problemas de conexión DeepSeek bloquean corridas	M	A	Retry con backoff; budget de contingencia
Falta tiempo para paper completo	B	M	Draft minimalista; secciones priorizadas
Taxonomía de errores insuficiente	B	B	50 ejemplos por dataset es mínimo viable

---

Decisiones de Descarte Documentadas

Este archivo **reemplaza** a `Sprints.md` para los sprints 3 y 4. Las tareas eliminadas quedan documentadas aquí para referencia pero **no se ejecutarán** dada la restricción temporal.

Filosofía: Entregar un proyecto **honesto, reproducible y bien documentado** aunque los resultados no sean espectaculares. Mejor un paper sólido diciendo "el método no mejoró baseline significativamente en estos datasets" que un proyecto incompleto.

Fecha de creación: 1 dic 2025
Última actualización: 1 dic 2025
Aprobado por: This au Chocolat (SM)