

Protocolo de Experimentación v2.0

Protocolo de Experimentación v2.0

Power10 + MMA: Benchmark de Anonimización Clínica

Universidad de Montevideo - Tesis 2025 Fecha: 19 de Diciembre 2025 Autor: Guillermo Robatto Para validación de: Tutor de Tesis

Resumen de Cambios vs. Versión Anterior

Aspecto	Versión Anterior (v1)	Nueva Versión (v2)
Casos de prueba	1 (Olaf Rasmussen)	3 (Caso A, Caso B, Olaf)
Prompts	1 fijo	6 estrategias diferentes
Métricas	Solo TPS	TPS + Precision + Recall + TP/FN
Evaluación	Manual	Automática con ground truth
Comparativa GPU	No existía	Tabla documentada con fuentes

Objetivos del Experimento

Objetivo Principal

Demostrar que IBM Power10 con aceleradores MMA puede ejecutar inferencia de LLMs para anonimización de datos clínicos con rendimiento y calidad suficientes para uso en producción.

Objetivos Secundarios

- Benchmark cuantitativo:** Medir tokens/segundo (TPS) en diferentes modelos
- Validación cualitativa:** Evaluar precision/recall de la anonimización
- Comparativa de prompts:** Determinar la mejor estrategia de prompting
- Comparativa teórica GPU:** Contextualizar rendimiento vs. alternativas cloud

Metodología

Dataset de Prueba

Se utilizan 3 casos clínicos sintéticos con ground truth anotado:

Caso	Descripción	Entidades PHI	Caracteres
caso_a	Emergencia Cardiología	10	~650
caso_b	Oncología Evolución	10	~550
caso.olaf	CTI Completo (multi-evolución)	20	~3,500

Archivo: benchmarks/casos_sinteticos.py

Ejemplo - Caso A (Emergencia Cardiología)

Paciente: Roberto Carlos Méndez, CI 3.456.789-1.
Consulta en Emergencia del Hospital Español el 14/11/2023.
Refiere dolor opresivo retroesternal...
Vive en Bulevar Artigas 2345, Montevideo.
Se llama a su hija Laura al 099-111-222...
Antecedentes: IAM previo tratado por Dr. Sanguinetti.

Entidades anotadas (ground truth): - [NOMBRE] Roberto Carlos Méndez - [CI] 3.456.789-1 - [UBICACION] Hospital Español - [FECHA] 14/11/2023 - [DIRECCION] Bulevar Artigas 2345 - [UBICACION] Montevideo - [NOMBRE] Laura - [TELEFONO] 099-111-222 - [NOMBRE] Dr. Sanguinetti

Estrategias de Prompting

Se comparan 6 estrategias diferentes:

ID	Nombre	Descripción	Tokens Est.
baseline	Prompt Simple	Instrucción mínima sin ejemplos	~80
detailed	Prompt Detallado	8 reglas específicas (original del proyecto)	~350
few_shot	Few-Shot Learning	3 ejemplos antes del texto	~300
chain_of_thought	Chain of Thought	Razonamiento en 5 pasos	~250
master_tutor	Master Prompt	Prompt del protocolo del tutor	~120
medico	Médico Especializado	Enfoque en Ley 18.331 y PHI	~280

Archivo: benchmarks/prompts_anonimizacion.py

Métricas de Evaluación

Métricas de Rendimiento

- **TPS (Tokens Per Second):** Velocidad de generación
- **Tiempo total (ms):** Latencia end-to-end
- **Desviación estándar:** Consistencia del rendimiento

Métricas de Calidad (Anonimización)

- **True Positives (TP):** Entidades correctamente anonimizadas
- **False Negatives (FN):** Entidades que se escaparon (no anonimizadas)
- **Precision:** $TP / \text{Total entidades} \times 100$
- **Recall:** $TP / \text{Total entidades} \times 100$
- **Entidades escapadas:** Lista detallada de fallos

Fórmula de evaluación:

```
for entidad in ground_truth:  
    if entidad.valor in texto_anonimizado:  
        FN += 1 # Se escapó  
    else:  
        TP += 1 # Fue reemplazada correctamente
```

Experimentos a Ejecutar

Experimento 1: Benchmark de Modelos

Objetivo: Comparar rendimiento de diferentes modelos LLM

Modelo	Puerto	Cuantización	RAM Estimada
Qwen2.5-7B	8089	Q4_K_M	~8 GB
Mistral-7B	8088	Q4_K_S	~6 GB
Phi-3.5-mini	8093	Q4_K_M	~3 GB
BioMistral-7B	8092	Q4_K_M	~5 GB
Mistral-Nemo-12B	8097	Q4_K_M	~8 GB

Comando:

```
python benchmark_anon.py --port <PUERTO> --caso caso_olaf --  
prompt detailed --iterations 5 --save
```

Output esperado:

RESULTADOS DEL BENCHMARK

TPS Promedio: 13.12 tokens/seg
Tiempo Promedio: 30500 ms

MÉTRICAS DE ANONIMIZACIÓN

Entidades PHI totales: 20
True Positives (TP): 19
False Negatives (FN): 1
Precision: 95.0%
Recall: 95.0%

⚠ ENTIDADES QUE SE ESCAPARON:
[NOMBRE] "Bremmerman" (licenciado enfermería)

Experimento 2: Comparativa de Prompts

Objetivo: Determinar la mejor estrategia de prompting

Comando:

```
python benchmark_prompts.py --caso caso_a --port 8089 --  
iterations 3 --export csv
```

Output esperado:

RESULTADOS COMPARATIVOS

# Prompt	TPS	Tiempo	TP/Total
Precision			
-----	-----	-----	-----
★1 Few-Shot Learning 100.0%	12.5	2500ms	10/10
2 Prompt Detallado 90.0%	13.1	2400ms	9/10
3 Master Prompt (Tutor) 90.0%	14.2	2200ms	9/10
4 Médico Especializado 80.0%	11.8	2600ms	8/10
5 Chain of Thought 80.0%	10.5	3000ms	8/10
6 Prompt Simple 60.0%	15.0	2100ms	6/10

Experimento 3: Matriz Completa

Objetivo: Evaluar todas las combinaciones caso × prompt × modelo

Comando:

```
python benchmark_anon.py --caso todos --prompt todos --port 8089  
--iterations 3 --save
```

Genera **18 combinaciones** (3 casos × 6 prompts) con resultados guardados en JSON.

Comparativa Teórica con GPUs

Archivo: docs/07-comparativa-gpu.md

Plataforma	TPS	Costo/hora	Privacidad
IBM Power10 (MMA)	13-20	\$0 (on-prem)	✓ Total
Nvidia T4 (Cloud)	20-25	~\$0.35	⚠ Cloud
Nvidia A10 (Cloud)	40-50	~\$0.80	⚠ Cloud
RTX 4090 (Consumer)	100-150	N/A	✓ Local

Argumento clave: > “Aunque GPUs cloud son 2-5x más rápidas, Power10 on-premises elimina > el riesgo de fuga de datos PHI y cumple automáticamente con Ley 18.331.”

Estructura de Archivos

```
Tesis-inco-grobatto/  
|   benchmarks/  
|   |   casos_sinteticos.py      # [NUEVO] Dataset con ground  
|   |   truth  
|   |   prompts_anonimizacion.py # [NUEVO] 6 estrategias de prompt  
|   |   benchmark_anon.py        # [MODIFICADO] Soporta --caso, --  
|   |   prompt  
|   |   benchmark_prompts.py    # [NUEVO] Comparador de prompts  
|   |   results/                 # Resultados JSON  
|   docs/  
|   |   07-comparativa-gpu.md    # [NUEVO] Comparativa teórica  
|   |   08-protocolo-experimento-v2.md # [NUEVO] Este documento  
|   ...
```

Cronograma de Ejecución

Stage 1: Preparación (COMPLETADO ✓)

- Crear dataset sintético con ground truth
- Implementar suite de 6 prompts
- Modificar benchmark para evaluación automática
- Crear script de comparación de prompts
- Documentar comparativa GPU

Stage 2: Ejecución en Power10 (PENDIENTE)

- Conectar al servidor Power10
- Verificar que MMA está activo
- Descargar modelos faltantes (Phi-3.5, BioMistral, Mistral-Nemo)
- Ejecutar Experimento 1 (benchmark modelos)
- Ejecutar Experimento 2 (comparativa prompts)
- Ejecutar Experimento 3 (matriz completa)
- Capturar screenshots de htop + terminal
- Generar tablas finales para la tesis

Entregables Esperados

1. **Tabla de rendimiento por modelo** (TPS, tiempo, RAM)
 2. **Tabla de calidad por prompt** (precision, recall, TP/FN)
 3. **Matriz caso × prompt × modelo** (resultados JSON)
 4. **Screenshots de evidencia** (htop, terminal con benchmarks)
 5. **Comparativa GPU documentada** (con fuentes académicas)
-

Preguntas para el Tutor

1. ¿Los 3 casos de prueba son suficientes o necesitamos más variedad?
 2. ¿Las 6 estrategias de prompting cubren las variantes relevantes?
 3. ¿La métrica de precision/recall basada en ground truth es apropiada?
 4. ¿Hay modelos adicionales que debamos probar (ej. Llama-3.1)?
 5. ¿Necesitamos incluir métricas de uso de memoria/CPU además de TPS?
-

Referencias

- [llama.cpp GitHub](#)
- [Baseten: Guide to LLM Inference](#)
- [LLM-Anonymizer NEJM AI](#)
- [Ley 18.331 Uruguay](#)