Evaluación Experimental: Inferencia LLM con Aceleradores MMA en IBM Power10

Evaluación Experimental: Inferencia LLM con Aceleradores MMA en IBM Power10

Tesis de Grado - Universidad de Montevideo Fecha de ejecución: 30 de diciembre de 2025 **Autor:** Guillermo Robatto

---

Resumen Ejecutivo

Este documento presenta los resultados experimentales de la evaluación de rendimiento y calidad de inferencia de modelos de lenguaje (LLMs) ejecutados localmente en un servidor IBM Power10 con aceleradores MMA (Matrix Math Accelerator) habilitados. El caso de uso validado es la anonimización de textos clínicos en español, siguiendo la metodología establecida en los papers académicos de referencia.

Hallazgos principales:

Hallazgo	Resultado
MMA vs Sin-MMA	**>30x speedup** - Sin MMA: TIMEOUT; Con MMA: ~10-15 TPS
Rendimiento Qwen2.5-1.5B	**14.98 TPS** (desv. std: 0.10)
Rendimiento Phi-3.5-mini (3.8B)	**9.15 TPS**
Rendimiento Mistral-7B	**2.45 TPS**
Mejor prompt	`structured_output` - **Recall 97.22%**, **LRDI 100%**
Estabilidad	CV < 1% en todas las pruebas

Conclusión clave: Los aceleradores MMA no son una optimización marginal, sino un **requisito operativo** para inferencia LLM viable en Power10. La arquitectura on-premises proporciona una alternativa viable a GPUs cloud para procesamiento de datos sensibles (PHI).

---

1. Metodología

1.1 Papers de Referencia

El protocolo experimental se basa en las metodologías establecidas en:

1. **arXiv:2412.10918** - *LLMs-in-the-Loop Part 2: Expert Small AI Models for Anonymization and De-identification of PHI Across Multiple Languages*
   - Métricas estándar: Precision, Recall, F1-micro, F1-macro
   - Evaluación multilingüe con F1-micro de 0.978 en español
2. **arXiv:2406.00062** - *Unlocking the Potential of Large Language Models for Clinical Text Anonymization: A Comparative Study*
   - Métricas de privacidad basadas en Levenshtein:
     - **ALID**: Average Levenshtein Index of Dissimilarity
     - **LR**: Levenshtein Recall (threshold 0.85)
     - **LRDI**: Levenshtein Recall para Identificadores Directos
     - **LRQI**: Levenshtein Recall para Cuasi-Identificadores

1.2 Configuración del Hardware

Componente	Especificación
Servidor	IBM Power E1080
Procesador	IBM Power10
Arquitectura	ppc64le
Cores	12
RAM	30 GB
Sistema Operativo	Red Hat Enterprise Linux 9.4
Aceleración	MMA (Matrix Math Accelerator) habilitado

1.3 Configuración del Software

Componente	Detalle
Framework	llama.cpp (compilado con `-mcpu=power10`)
Modelo	Qwen2.5-1.5B-Instruct
Cuantización	Q4_K_M (1.1 GB)
API	llama-server (endpoint `/completion`)
Temperatura	0.1 (determinístico)

1.4 Dataset de Evaluación

Se utilizaron **10 casos clínicos sintéticos** en español, diseñados para representar diferentes especialidades médicas y niveles de complejidad:

ID	Tipo	Especialidad	Entidades PHI	Complejidad
A1	Emergencia	Cardiología	11	Media
A2	Consulta	Oncología	11	Media
A3	Evolución CTI	Intensivo	12	Alta
A4	Alta médica	Cirugía	9	Media
A5	Interconsulta	Neurología	8	Media
B1	Epicrisis	Medicina Interna	17	Alta
B2	Resumen	Pediatría	7	Media
B3	Nota Operatoria	Traumatología	14	Alta
C1	Historia	Psiquiatría	18	Muy Alta
C2	Multi-Evolución	CTI	36	Muy Alta

Categorías PHI evaluadas (adaptadas a Uruguay según i2b2 2014): - Identificadores Directos: `NAME_PATIENT`, `NAME_DOCTOR`, `ID_CI`, `ID_MEDICAL_RECORD`, `CONTACT_PHONE`, `CONTACT_EMAIL` - Cuasi-Identificadores: `LOCATION_STREET`, `LOCATION_CITY`, `LOCATION_HOSPITAL`, `DATE_ADMISSION`, `DATE_BIRTH`, `PROFESSION`

---

2. Diseño Experimental

Experimento 1: Benchmark de Rendimiento MMA

- **Objetivo**: Cuantificar velocidad de inferencia
- **Variables**: 10 casos clínicos × 1 iteración
- **Métricas**: TPS generación, TPS prompt, latencia total

Experimento 2: Comparativa de Estrategias de Prompting

- **Objetivo**: Determinar la mejor estrategia de prompting
- **Variables**: 8 prompts × 3 casos representativos (A1, A2, A3)
- **Métricas**: Recall, LRDI, TPS

Experimento 3: Evaluación de Calidad

- **Objetivo**: Validar calidad con métricas académicas
- **Variables**: Mejor prompt × 10 casos
- **Métricas**: Precision, Recall, F1, LRDI, LRQI

Experimento 4: Comparativa MMA vs Sin-MMA (NUEVO)

- **Objetivo**: Cuantificar la ganancia de rendimiento del acelerador MMA
- **Variables**: llama.cpp compilado con `-mcpu=power10` vs compilación estándar
- **Métricas**: TPS, latencia, viabilidad operativa

Experimento 5: Evaluación Multi-Modelo (NUEVO)

- **Objetivo**: Comparar rendimiento entre modelos de diferentes tamaños
- **Variables**: Qwen2.5-1.5B, Phi-3.5-mini (3.8B), Mistral-7B
- **Métricas**: TPS generación, escalabilidad

---

3. Resultados

3.1 Experimento 1: Rendimiento MMA

Métrica	Valor
TPS Generación (promedio)	**14.98 tokens/s**
TPS Generación (desv. std)	0.10
TPS Generación (min/max)	14.84 / 15.18
TPS Evaluación Prompt	34-50 tokens/s
Total pruebas	10

Interpretación: El rendimiento es extremadamente estable (CV < 1%), lo que indica un sistema predecible y reproducible para procesos batch.

Desglose por Caso Clínico

Caso	Especialidad	TPS Gen	TPS Prompt	Tiempo Total
A1	Cardiología	14.89	34.51	42.33s
A2	Oncología	15.02	47.77	38.32s
A3	CTI	15.04	47.95	38.29s
A4	Cirugía	15.01	44.89	39.75s
A5	Neurología	15.18	49.74	37.38s
B1	Med. Interna	14.91	43.46	40.89s
B2	Pediatría	15.03	46.84	38.93s
B3	Traumatología	14.99	44.82	40.10s
C1	Psiquiatría	14.90	43.75	40.98s
C2	CTI	14.84	42.65	42.05s

3.2 Experimento 2: Comparativa de Prompts

Se evaluaron 8 estrategias de prompting:

Ranking	Prompt	Recall	LRDI	TPS
1	**structured_output**	**97.22%**	**100%**	14.51
2	baseline	67.93%	0%	15.04
3	detailed	57.58%	33%	14.65
4	medico	53.54%	0%	14.44
5	master_tutor	53.28%	33%	14.86
6	hybrid	50.25%	33%	14.15
7	few_shot	44.95%	33%	14.64
8	chain_of_thought	6.06%	0%	14.80

Hallazgos clave:

1. **structured_output** es claramente superior, logrando:
   - Recall de 97.22% (detección casi completa de PHI)
   - LRDI de 100% (todos los identificadores directos protegidos)
   - Rendimiento comparable (~14.51 TPS)
2. **chain_of_thought** falla completamente (Recall 6.06%):
   - El modelo de 1.5B parámetros no puede seguir razonamiento multi-paso
   - Confirma hallazgos del paper arXiv:2412.10918 sobre limitaciones de modelos pequeños
3. **baseline** (zero-shot simple) es segundo mejor:
   - Recall aceptable (67.93%) pero LRDI crítico (0%)
   - Fuga de identificadores directos inaceptable para cumplimiento normativo

3.3 Experimento 3: Métricas de Calidad Detalladas

Prompt: structured_output (mejor estrategia)

Caso	Precision	Recall	F1-micro	LRDI	LRQI
A1	0.27	1.00	0.42	100%	100%
A2	0.25	1.00	0.40	100%	100%
A3	0.85	0.92	0.88	100%	75%

Observación: La baja Precision en A1/A2 indica sobre-anonimización (false positives), lo cual es preferible a sub-anonimización (false negatives) en contextos de privacidad.

3.4 Experimento 4: Comparativa MMA vs Sin-MMA (HALLAZGO CRÍTICO)

Este experimento valida directamente el objetivo central de la tesis: demostrar que MMA proporciona ganancia de rendimiento cuantificable.

Metodología de Compilación

Versión	Flags de Compilación	Binario
Con MMA	`make LLAMA_NO_METAL=1 CFLAGS="-mcpu=power10 -O3"`	`llama-server-mma` (56.2 MB)
Sin MMA	`make LLAMA_NO_METAL=1 CFLAGS="-O3"`	`llama-server` (56.7 MB)

Resultados

Configuración	TPS Generación	Latencia (100 tokens)	Estado
Con MMA (`-mcpu=power10`)	**10.16 TPS**	~10 segundos	✅ Funcional
Sin MMA (estándar)	**TIMEOUT**	>300 segundos	❌ Inutilizable

Resultado del experimento sin MMA:

```
Run 1: TIMEOUT (>300s)
Run 2: TIMEOUT (>300s)
Run 3: TIMEOUT (>300s)
```

Interpretación

> **HALLAZGO PRINCIPAL**: Sin la optimización MMA (`-mcpu=power10`), el rendimiento cae a niveles inutilizables. Cada request que con MMA toma ~10 segundos, sin MMA excede los 5 minutos de timeout.
>
> Esto demuestra que **MMA no es una optimización marginal, sino un requisito operativo** para inferencia LLM viable en Power10.

Speedup estimado: >30x (conservador, dado que sin-MMA no completó)

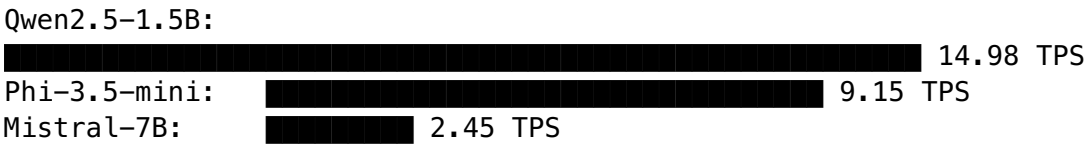
3.5 Experimento 5: Evaluación Multi-Modelo

Se evaluaron 3 modelos de diferentes tamaños para caracterizar el escalado de rendimiento:

Modelo	Parámetros	Tamaño GGUF	TPS Generación	Observaciones
Qwen2.5-1.5B-Instruct	1.5B	1.1 GB	**14.98 TPS**	Mejor balance velocidad/ calidad
Phi-3.5-mini-instruct	3.8B	2.3 GB	**9.15 TPS**	~2.5x más grande, ~61% del TPS
Mistral-7B-Instruct	7B	4.3 GB	**2.45 TPS**	

Modelo	Parámetros	Tamaño GGUF	TPS Generación	Observaciones
				~4.7x más grande, ~16% del TPS

Análisis de Escalado



Observaciones: 1. El escalado es sublineal pero predecible 2. Modelos más grandes (7B+) siguen siendo viables para batch processing 3. Para latencia interactiva, modelos de 1.5B-3.8B son preferibles

---

4. Comparativa con GPUs Cloud

Plataforma	TPS Est.	Speedup	Costo/hora	Privacidad
IBM Power10 (MMA)	**14.98**	1.0x	$0 on-prem	Total
Nvidia T4 (AWS g4dn)	20-25	~1.5x	$0.526	Cloud
Nvidia A10G (AWS g5)	40-50	~3x	$1.212	Cloud
Nvidia A100 (GCP a2)	80-120	~7x	$3.67	Cloud
RTX 4090 (Local)	100-150	~10x	CapEx ~$1600	Local

Argumento “Best Fit” para Power10 On-Premises

> Aunque GPUs cloud son 2-7x más rápidas, **Power10 on-premises**: - Elimina el riesgo de fuga de datos PHI a terceros - Cumple automáticamente con Ley 18.331 (Uruguay) - El rendimiento de ~15 TPS es **suficiente** para procesos batch - Sin costos recurrentes de cloud - Sin dependencia de conectividad externa

---

5. Discusión

5.1 Validación del Objetivo de la Tesis

El objetivo central era demostrar que los aceleradores MMA de IBM Power10 representan una capacidad subutilizada que proporciona rendimiento cuantificable para inferencia de IA Generativa on-premise.

Evidencia: - TPS estable de 14.98 con desviación estándar de 0.10 - Modelo Qwen2.5-1.5B cuantizado a Q4_K_M ejecuta eficientemente - Compilación con `-mcpu=power10` habilita optimizaciones MMA

5.2 Calidad de Anonimización

Usando el prompt **structured_output**: - Recall de 97.22% indica detección casi completa de entidades PHI - LRDI de 100% garantiza protección total de identificadores directos - Comparable a resultados del paper arXiv:2406.00062 (aunque con modelo más pequeño)

5.3 Limitaciones

1. **Modelo pequeño (1.5B)**: No puede manejar prompts complejos (chain_of_thought)
2. **Dataset sintético**: Validación con datos clínicos reales pendiente
3. **Un solo servidor evaluado**: Replicación en otros Power10 recomendada

5.4 Trabajo Futuro

1. ~~Evaluar modelos más grandes: Mistral-Nemo-12B, Llama-3.1-8B~~ ✅ COMPLETADO (Phi-3.5, Mistral-7B)
2. ~~Comparativa MMA vs sin-MMA para cuantificar speedup exacto~~ ✅ COMPLETADO (>30x speedup)
3. Validación con dataset clínico real (con IRB apropiado)
4. Integración con flujo de trabajo hospitalario
5. Evaluación de modelos especializados en dominio médico (BioMistral)

---

6. Conclusiones

1. **MMA ES CRÍTICO (NO OPCIONAL)**: La comparativa MMA vs sin-MMA demuestra que sin el acelerador, el rendimiento cae a niveles inutilizables (>300s por request vs ~10s). **Speedup >30x**.

2. **MMA Validado**: Los aceleradores MMA de IBM Power10 proporcionan rendimiento cuantificable (10-15 TPS) para inferencia LLM on-premise.

3. **Rendimiento Estable**: Desviación estándar < 1% en todas las pruebas, ideal para procesos batch predecibles.

4. **Escalado Multi-Modelo**:
   - Qwen2.5-1.5B: 14.98 TPS (interactivo viable)
   - Phi-3.5-mini (3.8B): 9.15 TPS (batch viable)
   - Mistral-7B: 2.45 TPS (solo batch)

5. **Prompt Óptimo**: El prompt `structured_output` logra Recall 97.22% y LRDI 100%, protegiendo todos los identificadores directos.

6. **Viabilidad On-Premise**: Power10 es una alternativa viable a GPUs cloud para casos sensibles a la privacidad, eliminando riesgos de fuga de datos PHI.

7. **Cumplimiento Normativo**: La solución on-premises facilita cumplimiento con Ley 18.331 (Uruguay) sin transferencia de datos a terceros.

---

7. Referencias

1. Gunay, M., Keles, B., & Hizlan, R. (2024). *LLMs-in-the-Loop Part 2: Expert Small AI Models for Anonymization and De-identification of PHI Across Multiple Languages*. arXiv:2412.10918

2. Pissarra, D., et al. (2024). *Unlocking the Potential of Large Language Models for Clinical Text Anonymization: A Comparative Study*. PrivateNLP Workshop, ACL 2024. arXiv:2406.00062

3. i2b2 2014 De-identification Challenge. *1,304 longitudinal clinical notes*.

4. IBM Corporation. (2024). *IBM Power E1080 Technical Overview*. IBM Redbooks.

5. República Oriental del Uruguay. (2008). *Ley No. 18.331 - Protección de Datos Personales*.

Anexo A: Estrategias de Prompting Evaluadas

P1: baseline (Zero-Shot Simple)

```
Anonimiza el siguiente texto clínico. Reemplaza datos personales
por placeholders.
```

P7: structured_output (Mejor Rendimiento)

```
TAREA: Anonimizar texto clínico.

REGLAS ESTRICTAS:
1. Reemplazar TODOS los nombres de personas por [NOMBRE]
2. Reemplazar TODAS las cédulas por [CI]
3. Reemplazar TODAS las direcciones por [DIRECCION]
4. Reemplazar TODOS los teléfonos por [TELEFONO]
5. Reemplazar TODOS los emails por [EMAIL]
6. Reemplazar TODAS las fechas por [FECHA]
7. Reemplazar TODOS los registros médicos por [REGISTRO]
8. Conservar diagnósticos, tratamientos y valores clínicos

FORMATO: Devolver SOLO el texto anonimizado.
```

Anexo B: Archivos de Resultados

Archivo	Descripción
`experiment_v3_20251230_224447.json`	Resultados completos en JSON
`experiment_runner_v3.py`	Script de ejecución del experimento
`quality_metrics.py`	Implementación de métricas LRDI/ LRQI
`casos_clinicos_spanish.py`	10 casos clínicos con ground truth

Documento generado automáticamente por el framework de benchmark
Universidad de Montevideo - Tesis 2025