Taxonomías de habilidades para la evaluación de sistemas inteligentes

Coloquio SADAF 2025

Mauro Santelli

Trabajo en colaboración con Joaquín Toranzo Calderón y Ramiro Caso

IIF-SADAF-CONICET & Universidad de Buenos Aires

21 de octubre de 2025

El problema: evaluar modelos de lenguaje

Contexto: GPT-3 y los grandes modelos de lenguaje

- Entrenados con cantidades masivas de datos
- Capacidad multi-tarea sin entrenamiento específico
- Propiedades emergentes no anticipadas

Problema

No tenemos marcos conceptuales adecuados para evaluar estas capacidades.

Preguntas que guían este trabajo

- ¿Qué miden nuestros tests?
- ¿Cómo se relacionan las capacidades evaluadas?
- ¿Qué significa que un modelo "comprenda" o "razone"?

Benchmarking: el enfoque dominante

Tipos de evaluación (Chollet, 2019)

- Evaluación humana (costosa, no escalable)
- Evaluación de caja blanca (requiere acceso interno)
- Evaluación adversarial (difícil de estandarizar)
- **Benchmarking** (dominante en la práctica)

Por qué dominan los benchmarks

Estandarización, comparabilidad, reproducibilidad, eficiencia computacional.

Definición

Benchmark = Conjunto de problemas + Función de puntuación

Ejemplo: dataset de pregunta-respuesta (QA)

Estructura típica

- Dataset: pares (pregunta, respuesta correcta)
- Ejemplo: "¿Cuál es la capital de Francia?" $\rightarrow$ "París"
- Dominios: conocimiento general, especializado, razonamiento

Respuestas correctas

- Únicas o múltiples
- Generadas por humanos o extraídas automáticamente

Problema

¿Qué habilidad específica estamos midiendo? No es obvio.

Funciones de puntuación

Exact Match (EM)

- Comparación exacta de cadenas
- Binaria: correcto o incorrecto
- Simple pero restrictiva

Métricas semánticas

- Similitud vía embeddings
- Permite variación expresiva
- Más flexible, requiere calibración

Otras métricas

BLEU, ROUGE (generación), F1 (extracción), Accuracy, Precision, Recall.

Observación

La elección de métrica determina qué cuenta como “éxito”.

Enfoques actuales de benchmarking

Progresión histórica

1. Datasets individuales (MNIST, SQuAD)
2. Conjuntos de benchmarks agrupados (GLUE, SuperGLUE)
3. Grandes suites (BIG-bench)
4. Suites + criterios de selección (HELM)

Ejemplos

- **Eleuther LM Harness**: framework técnico
- **HELM**: evaluación holística de Stanford

Tendencia

Proliferación de benchmarks sin marco unificador teórico.

HELM: Holistic Evaluation of Language Models

Características (Liang et al., 2022)

- Más de 50 escenarios de evaluación
- Dimensiones: precisión, calibración, robustez, justicia, eficiencia
- Documentación exhaustiva

Organización de tareas

- Categorías amplias: QA, Information Retrieval, Summarization, Reasoning
- Agrupaciones por similitud superficial
- Sin estructura jerárquica explícita

Admisión de los autores

"We do not claim to select scenarios based on first principles... [our] coverage is opportunistic rather than systematic."

Nuestra propuesta: Taxonomías de Habilidades Artificiales

Objetivo

Desarrollar un vocabulario estructurado sobre capacidades para organizar la evaluación.

Características

- Taxonomías jerárquicas
- Relaciones explícitas entre habilidades
- Fundamentación filosófica
- Compromisos teóricos mínimos

Pluralismo

- No imponer una sola teoría
- Permitir múltiples instanciaciones
- Hacer explícitos los compromisos

Beneficio

Reemplazar la selección ad hoc de benchmarks por decisiones teóricamente informadas.

El desafío del loro estocástico

"Un modelo de lenguaje es un sistema para unir secuencias de formas lingüísticas de manera azarosa... sin referencia al significado: un loro estocástico." (Bender et al., 2021)

Requisitos para competencia lingüística genuina

1. **Trasfondo común** (common ground) + conciencia situacional
2. **Intenciones comunicativas** genuinas
3. **Teoría de la mente** (atribución de estados mentales)
4. **Responsabilidad** (accountability) social

Problema

¿Cómo operacionalizamos estos requisitos? ¿Cómo los evaluamos?

Tradición pragmatista-cibernética

De Dewey a Brandom vía la cibernética

- Prioridad del conocimiento práctico (know-how) sobre el proposicional (know-that)
- Habilidades como bucles de retroalimentación
- Composicionalidad: habilidades complejas emergen de simples
- Realizabilidad múltiple: distintos caminos para la misma capacidad

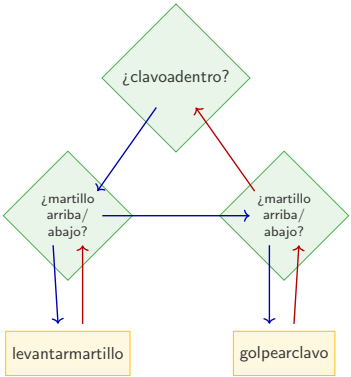
Herramientas conceptuales

- **MUDs** (Brandom): relaciones entre prácticas y vocabularios
- **TOTEs** (Miller et al., 1960): Test-Operate-Test-Exit

Tesis

Las habilidades pueden elaborarse como estructuras recursivas de prueba y operación.

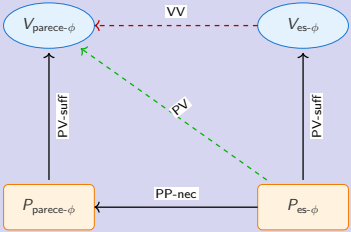
Ejemplo: Ciclos TOTE



Estructura

Diamantes = pruebas (condiciones); rectángulos = operaciones (acciones). Flechas azules = flujo; rojas = retroalimentación.

MUDs de Brandom: ejemplo Sellars



Interpretación

Rectángulos = prácticas (P); elipses = vocabularios (V). Relaciones: PV-suff (suficiencia), PP-nec (necesidad). Las relaciones práctica-vocabulario explican cómo emergen competencias complejas.

Chollet (2019): “On the measure of intelligence”

Crítica

- Evaluaciones “basadas en destrezas” y estrechas
- Tareas definidas de manera precisa
- No miden habilidades generalizables

Propuesta

- **Habilidad**: constructo abstracto
- **Destreza**: rendimiento en tarea específica
- **Inteligencia**: eficiencia en adquirir destrezas

Sobre generalización

- Destreza es específica a una tarea
- Inteligencia es capacidad meta de aprender nuevas destrezas
- Falta de generalización fuera de distribución: obstáculo central

Jerarquía de habilidades

Antecedente: psicometría

- Factor g (inteligencia general)
- Capacidades intermedias
- Tareas concretas en la base

Para sistemas artificiales

1. **Capacidad**: general
2. **Habilidad**: intermedia
3. **Destreza**: específica
4. **NTCS**: benchmark concreto

Distinción

Inteligencia = eficiencia en adquirir habilidades nuevas, no colección de destrezas.

Taxonomías de habilidades artificiales (TAAs)

Definición

Mapeo jerárquico: de lo general (capacidades) a lo específico (tareas).

Capacidad $\rightarrow$ **Habilidad** $\rightarrow$
Destreza $\rightarrow$ **NTCS**

Características

- **Jerárquico**: general arriba, específico abajo
- **Composicional**: complejas desde simples
- **Realizabilidad múltiple**: distintos caminos
- **Testeable**: relaciones permiten predicción

Nota

"Habilidades" = término genérico para cualquier nivel.

Tipos de relaciones entre habilidades

Relaciones metafísicas

- **Constitución**: B y C constituyen A
- **Superveniencia**: A superviene sobre {B C}
- **Gradación**: más de B mejora A

Relaciones lógicas

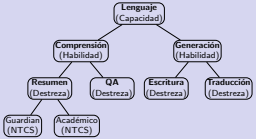
- **Necesidad**: B es necesaria para A
- **Suficiencia**: B es suficiente para A
- **Conjunción**: {B, C} juntas necesarias

Notación: $\triangleright$

A $\triangleright$ {B, C, D}: A requiere/se descompone en B, C, D.

Variantes: A $\triangleright_{\text{nec}}$ B (necesidad), A $\triangleright_{\text{suff}}$ B (suficiencia), A $\triangleright$ {B $\vee$ C} (realizabilidad múltiple).

Ejemplo: jerarquía de competencia lingüística



Cada nivel agrupa capacidades relacionadas. NTCS = benchmarks específicos.

El loro estocástico como TAA

Operacionalización de Bender et al.

Sea [LCC] = Competencia Comunicativa Lingüística. Entonces:

[LCC] ▷ {[CB], [CLI], [ToM], [Acc]}

CB = Trasfondo común (Common Background)

CLI = Intenciones comunicativas

ToM = Teoría de la mente

Acc = Responsabilidad (Accountability)

Descomposición recursiva

CB ▷ {conocimiento cultural, interpretación contextual}

CLI ▷ {planificación de objetivos, coherencia}

ToM ▷ {seguimiento de creencias, reconocimiento de intenciones}

TAA del loro estocástico: sub-habilidades testeables

[CB] Trasfondo

Sub-habilidades:

- Conocimiento cultural
- Interpretación contextual

Tests:

- Razonamiento contextual
- Referencias culturales

[CLI] Intenciones

Sub-habilidades:

- Planificación de objetivos
- Coherencia discursiva

Tests:

- Diálogo dirigido
- Comunicación estratégica

[ToM] Teoría de la mente

Sub-habilidades:

- Seguimiento de creencias
- Toma de perspectiva

Tests:

- Tareas de falsa creencia
- Perspectiva múltiple

Resultado

Los requisitos filosóficos de Bender et al. se vuelven empíricamente testeables.

Ejemplo: NLI como tarea a descomponer

Natural Language Inference (NLI)

Determinar si una hipótesis se sigue de una premisa (implicación, contradicción, neutral).

Descomposición en sub-habilidades

[NLI] $\triangleright$ {[Lex], [Gram], [LogInf], [MonInf], [PragRes]}

[Lex] = léxico, [Gram] = gramática, [LogInf] = inferencia lógica, [MonInf] = inferencia monotónica, [PragRes] = pragmática.

Implicación

SNLI, MultiNLI, ANLI evalúan NLI pero enfatizan distintas sub-habilidades.

Datasets NLI: análisis comparativo

SNLI

- Énfasis: [Lex], [Gram]
- Inferencias simples
- 570k ejemplos

MultiNLI

- Énfasis: [MonInf]
- Múltiples géneros
- 433k ejemplos

ANLI

- Énfasis: [LogInf], [PragRes]
- Adversarial
- 163k ejemplos

Observación

Una TAA permite entender qué evalúa cada dataset y por qué un modelo puede funcionar bien en uno pero mal en otro.

De evaluación plana a estructurada

Enfoque actual

- Lista plana de benchmarks
- Agrupaciones superficiales
- Sin estructura explicativa
- Selección ad hoc

Enfoque TAA

- Selección con principios
- Jerarquía explicativa
- Predicción entre tareas
- Capacidad diagnóstica

Ejemplo diagnóstico

Si un modelo falla en [Resumen] y [QA], la TAA sugiere inspeccionar [Comprensión de Texto]. Si [Comprensión] es débil, enfocar entrenamiento ahí; si es fuerte, buscar otros cuellos de botella.

Pluralismo metodológico

Por qué evitar compromisos fuertes

- Múltiples perspectivas contribuyen
- Evitar bloqueos teóricos
- Colaboración interdisciplinaria

Pluralismo como estrategia

- Distintas TAAs desde distintas teorías
- Predicciones empíricas comparan teorías
- TAA = esqueleto común

Ejemplo: "comprensión genuina"

Funcionalistas (mapeos E-S), mecanicistas (procesos internos), pragmatistas (éxito contextual) generan TAAs con predicciones distintas.

Perspectivas filosóficas compatibles

Compromisos teóricos

- **Funcionalistas**: mapeos E-S
- **Mecanicistas**: procesos internos
- **Pragmatistas**: éxito contextual
- **Inferencialistas**: inferencias

Beneficios

- Mismo esqueleto, distintas instanciaciones
- Explicita divergencias
- Disputas resolubles empíricamente

Punto central

Las TAAs ofrecen un lenguaje común para expresar posiciones y someterlas a prueba.

Conclusiones

Recapitulación

1. **Problema**: evaluación de LLMs sin estructura con principios
2. **Recursos**: filosofía de habilidades (pragmatismo, psicometría)
3. **Marco**: TAAs combinan jerarquía con flexibilidad teórica
4. **Beneficios**: vocabulario claro, capacidad diagnóstica, colaboración
5. **Impacto**: debates abstractos se vuelven testeables

Tesis

No es necesario resolver todos los problemas filosóficos para avanzar: hacen falta marcos que expliciten y permitan testear compromisos teóricos.

Agenda de investigación

Desarrollo técnico

1. TAAs para capacidades centrales (lenguaje, razonamiento)
2. Mapear benchmarks existentes a jerarquías
3. Evaluar poder predictivo de las TAAs
4. Herramientas diagnósticas

Colaboración

- Diálogo con HELM y otros marcos
- Vocabulario compartido IA/filosofía
- Marcos abiertos a múltiples teorías

Próximo paso

TAAs concretas para dominios prioritarios, comenzando con competencia lingüística.

Por qué importa ahora

Para IA

- Del "benchmark climbing" a evaluación con principios
- Entender qué medimos
- Base para discutir AGI

Para filosofía

- Dominio concreto para teorías de habilidades
- Banco de pruebas empírico
- Puente teoría-aplicación

Para colaboración

- Vocabulario común
- Compromisos explícitos
- Ruta estructurada

Contexto

Los LLMs se despliegan masivamente sin comprensión clara de sus capacidades. Mejores marcos evaluativos son urgentes.

Bibliografía

Referencias

- Bender, E. M., et al. (2021). "On the Dangers of Stochastic Parrots." *FAccT*.
- Brandom, R. (1994). *Making It Explicit*. Harvard UP.
- Chollet, F. (2019). "On the Measure of Intelligence." *arXiv:1911.01547*.
- Liang, P., et al. (2022). "Holistic Evaluation of Language Models." *arXiv:2211.09110*.
- Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the Structure of Behavior*. Holt.
- Ryle, G. (1949). *The Concept of Mind*. U. Chicago Press.
- Sellars, W. (1956). "Empiricism and the Philosophy of Mind." *MSPS*.
- Stanley, J. (2011). *Know How*. Oxford UP.

Discusión

Preguntas abiertas

- ¿Qué capacidades priorizar para las primeras TAAs?
- ¿Cómo equilibrar riqueza teórica con usabilidad práctica?
- ¿Cómo validar empíricamente las predicciones TAA?

Gracias

Mauro Santelli · IIF-SADAF-CONICET & UBA
En colaboración con Joaquín Toranzo Calderón y Ramiro Caso