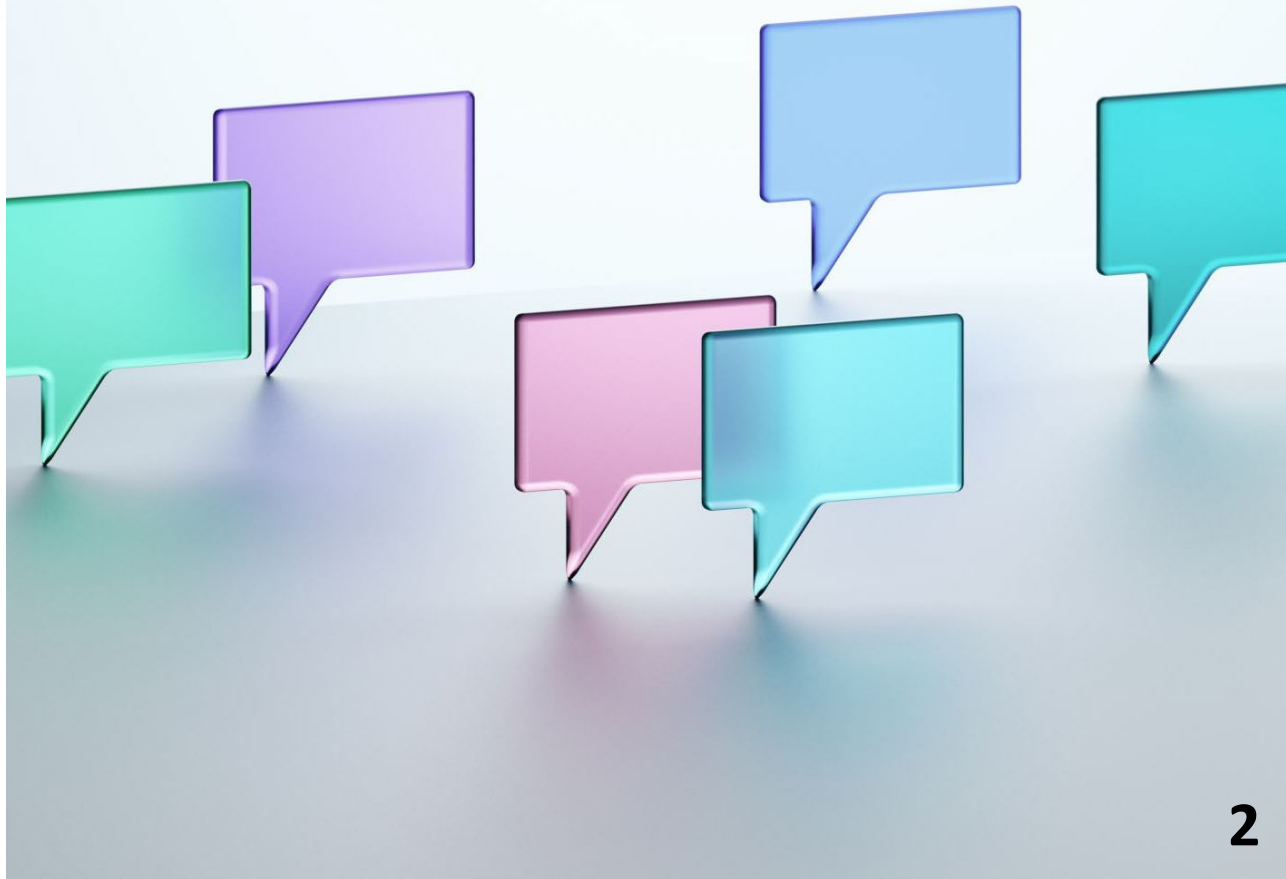
Evaluación automática de sistemas de diálogo mediante LLMs y generación de recursos en castellano

Luis Fernando D'Haro (@lfdharo)
Grupo de Tecnología del Habla y Aprendizaje Automático
ETSI de Telecomunicación - Universidad Politécnica de Madrid

I Workshop CLARIAH-CM: Humanidades Digitales y Tecnologías del Lenguaje
Madrid, mayo 30 de 2024

Contenido

- Motivación
- Evaluación automática de sistemas de diálogo con LLMs
- Recursos en castellano
- Conclusiones



Motivación

- Sistemas actuales como ChatGPT, Llama3, Gémini o Claude nos sorprenden por sus capacidades generativas (incluso en el tratamiento de diálogos)

- El entrenamiento de estos sistemas requiere cantidades muy grandes de datos para el pre-entrenamiento y de instrucciones para su alineamiento.

- Sin embargo, la evaluación humana (incluso ChatbotArena) es costosa y ruidosa, y la evaluación automática aún dista de ser fácil y confiable
  - Debemos añadir la dificultad de utilizar múltiples lenguas (escasez de datos + pocos LLMs multilingües)

Evaluación Automática

Evaluación Humana

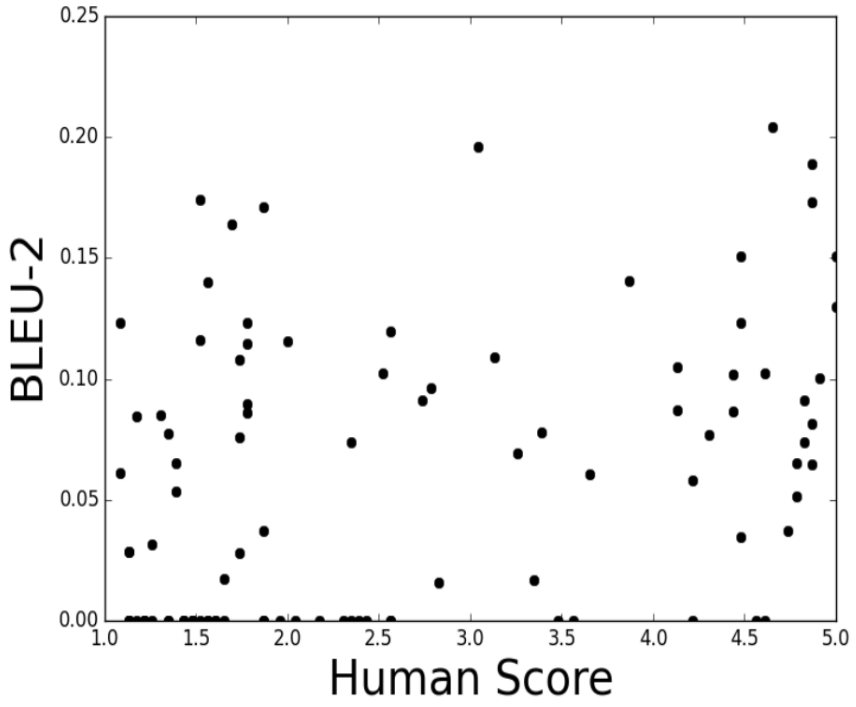
- Es la opción por defecto para cuantificar el progreso en las capacidades generativas de los LLMs. Características:
  - **Niveles**: turno y diálogo.
  - **Precisión**: Tenemos la capacidad de entender el lenguaje natural.
  - **Multidimensional**: Tenemos la capacidad de juzgar la generación integrando diferentes perspectivas.
  - **Acuerdo**: Se usa la votación por mayoría entre multiples anotadores.
- Pero…
  - Hay falta de consistencia (ej., expertos vs no expertos vs trabajadores colaborativos)
  - Depende de la edad, el estado de ánimo, la cultura, el conocimiento del tema, la experiencia previa, las expectativas,...
  - Son costosos y requieren mucho tiempo, especialmente si se aumenta el número de dimensiones a evaluar.
  - Están muy condicionados por la configuración de la evaluación (es decir, la escala, la descripción de la tarea, la selección de anotadores, el control de calidad, etc.).

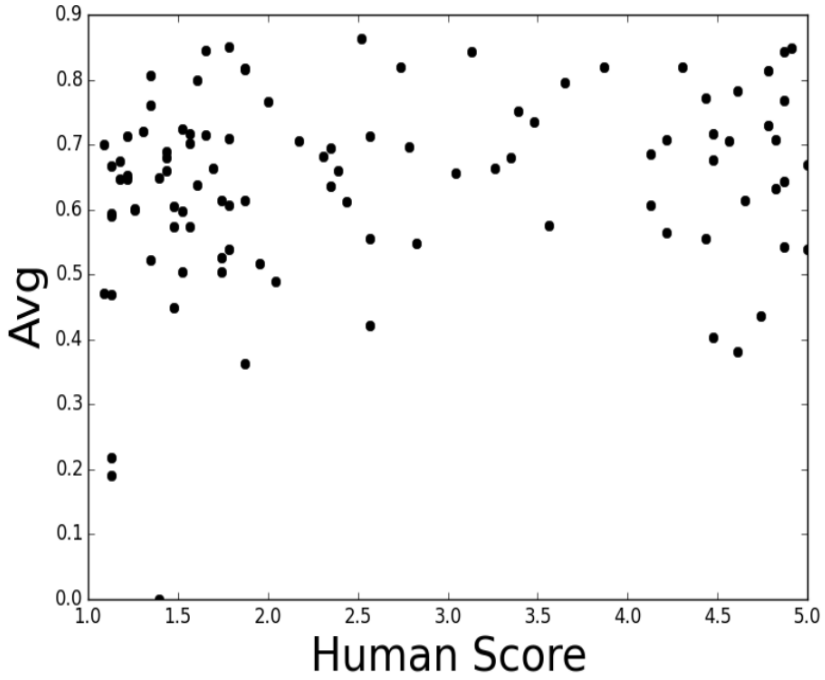
Evaluación automática

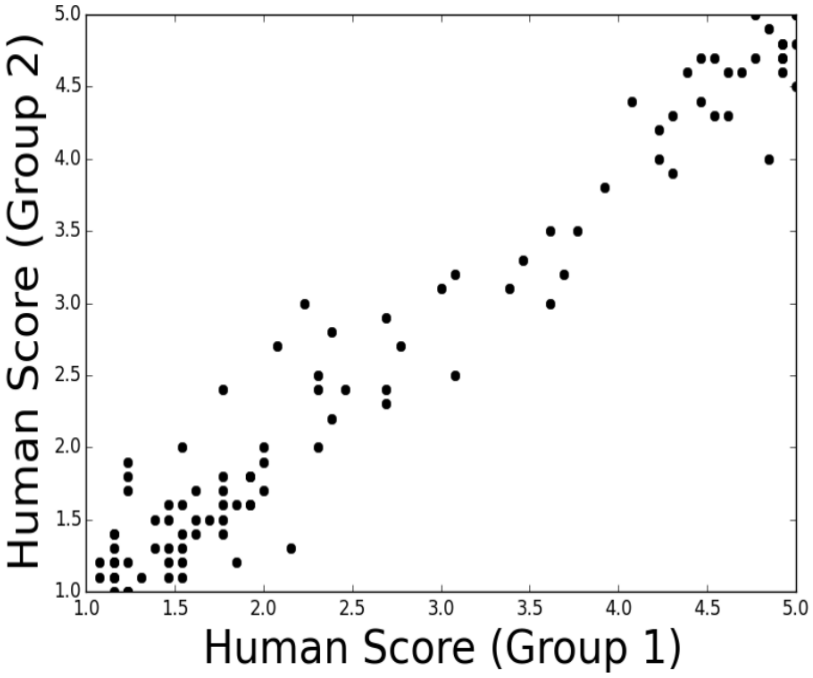
- Su objetivo no es reemplazar la evaluación humana, sino complementarla con una evaluación que sea consistente, reproducible, eficiente y económica.
- Atributos favorables de las métricas automáticas
  - Fuerte correlación con el juicio humano.
  - Interpretabilidad y multidimensionalidad.
  - Generalizable a través de diferentes dominios.
  - Robustez frente a ataques adversarios.
  - Compatible con la evaluación humana.
- En contra:
  - Falta de bases de datos, variabilidad en las anotaciones, obsolescencia de los sistemas, multiplicidad de respuestas (no se puede usar siempre la misma referencia).
  - Dificultad para integrar múltiples factores de evaluación, reducida capacidad para considerar el sentido común, multilingüalidad, multiculturalidad, etc.

Métricas típicas y sus problemas

- Las métricas más comunes son: BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), etc.
- Escasa correlación con la evaluación humana (Liu et al., 2016)
- No tienen en cuenta la variabilidad semántica + personalización







- Participación en el Amazon Alexa Prize Socialbot Grand Challenge (SGC5)
- Duración: 14 de noviembre de 2022 - 31 de agosto de 2023.
- 9 equipos seleccionados de todo el mundo: 7 universidades americanas (incluyendo Stanford, Universidad de California y Carnegie Mellon) + 2 europeas (R. Checa y UPM).
- Uso de nuestras métricas estado de la cuestión: baja correlación con puntuaciones humanas (0.2 – 0.3)

Proyecto Europeo ASTOUND

Funded by the European Union

ASTOUND: Mejora de las competencias sociales de agentes virtuales a través de una conciencia artificial basada en la Teoría del Esquema de la Atención.

Objetivos:

- Incorporación de nuevos mecanismos de controlabilidad en chatbots conversacionales.
- Gráficos de conocimiento + automatización del perfil de la persona + resumen de diálogo.
- Conciencia contextual a través de la integración de múltiples modalidades (voz, video, texto, wearables)
- Nuevas métricas de autoevaluación para nuevas dimensiones: interacción a largo plazo, memoria, sesgo, toxicidad.

- **Fechas: Dic 2022 – Nov 2025**
- **Participantes:**
- Ecole Normale Superieure (Francia), University Medical Center Hamburg-Eppendorf (Alemania), Alien Technology Transfer (Italia)
  - Princeton University (USA), Microsoft y MILA (Canadá)

Creación de bbdd multilingüe + uso de LLMs

- xDial-Eval*: Traducción de +18 bb.dd con 14,930 turnos anotados y 8,691 diálogos a 10 idiomas.
  - Chino, Español, Alemán, Francés, Japonés, Coreano, Ruso, Indio y Árabe.
- Tanto la evaluación automática de la traducción como la humana validan la alta calidad de xDial-Eval.
  - Servicio de MT de Microsoft Azure (costo ~ 400 USD)

Turn-Level Datasets	#Instance	#Utts/Instance	#Ctx/Hyp Words	#Dims
Persona-USR (2020b)	300	9.3	98.4 / 12.0	6
Persona-Zhao (2020)	900	5.1	48.8 / 11.5	4
ConvAI2-GRADE (2020)	600	3.0	24.4 / 11.3	1
Persona-DSTC10 (2022b)	4,829	4.0	36.0 / 11.6	4
DailyDialog-GRADE (2020)	300	3.0	26.0 / 10.8	1
DailyDialog-Zhao (2020)	900	4.7	47.5 / 11.0	4
DailyDialog-Gupta (2019)	500	4.9	49.9 / 10.9	1
Topical-USR (2020b)	360	11.2	236.3 / 22.4	6
Topical-DSTC10 (2022b)	4,500	4.0	50.6 / 15.9	4
Empathetic-GRADE (2020)	300	3.0	29.0 / 15.6	1
FED-Turn (2020a)	375	10.4	87.3 / 13.3	9
ConTurE-Turn (2022a)	1066	3.8	21.67 / 10.99	1
Dialogue-Level Datasets	**#Instance**	**#Utts/Instance**	**#Words/Utt**	**#Dims**
IEval (2022)	1,920	6.0	12.4	8
Persona-See (2019)	3,316	12.0	7.6	9
Reliable-Eval (2022)	2,925	21.2	8.4	7
ConTurE-Dial (2022b)	119	17.9	8.6	11
FED-Dial (2020a)	125	12.7	9.2	11
Human-Eval (2022)	286	12.0	11.6	3

* Chen Zhang, **Luis F. D'Haro**, Chengguang Tang, Ke Shi, Guohua Tang, and Haizhou Li. 2023. xDial-Eval: A Multilingual Open-Domain Dialogue Evaluation Benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5579–5601, Singapore. Association for Computational Linguistics.

Calidad de la traducción de los datos

- 350 pares de traducción por anotador (5) por cada idioma, calificándolos a mano en una escala del 1 al 5, calificación humana promedio ~ 4.5

Lang	BLEU	BERTScore	BLEURT	GPT-4	CometKiwi22	CometKiwi23
EN-ZH	37.85	0.691	0.773	4.58	0.827	0.739
EN-ES	56.58	0.767	0.820	4.64	0.845	0.764
EN-DE	50.30	0.754	0.817	4.63	0.834	0.744
EN-FR	52.10	0.748	0.811	4.59	0.842	0.709
EN-JA	42.70	0.711	0.786	4.33	0.854	0.778
EN-KO	40.23	0.707	0.782	4.26	0.846	0.767
EN-HI	49.33	0.726	0.797	4.40	0.793	0.699
EN-AR	48.54	0.737	0.794	4.39	0.835	0.736
EN-RU	49.13	0.741	0.804	4.40	0.825	0.739
Avg	47.42	0.731	0.798	4.47	0.833	0.742

Resultados a nivel de turno usando LLMs

- Múltiples LLMs (abiertos y cerrados) con y sin ajuste fino.
- Combinación con métrica de referencia previa (POE*)
- ChatGPT muy cercano a POE, aunque el ajuste fino de los LLMs permite obtener resultados mejores.
- Combinación de modelos permite mejoras, pero no significativas

		Turn-Level										
Category	**Models**	**EN**	**ZH**	**ES**	**DE**	**FR**	**JA**	**KO**	**HI**	**AR**	**RU**	**AVG**
BERT-Based	PoE†	0.464	0.437	0.441	0.454	0.455	0.424	0.417	0.361	0.422	0.436	0.431
LLMs-Zeroshot	LLaMA-7B	0.038	0.025	0.094	0.028	0.037	0.071	0.015	-0.020	0.016	0.072	0.038
	LLaMA-2-7B	0.065	0.076	0.084	0.029	0.033	0.101	0.108	0.066	0.073	0.010	0.064
	BLOOM-7B	0.044	0.134	0.100	0.019	0.084	0.017	0.005	0.048	0.099	0.062	0.061
	Falcon-7B	0.143	0.127	0.155	0.088	0.151	0.093	0.011	0.068	0.109	0.077	0.102
	Baichuan-2-7B	0.175	0.134	0.118	0.133	0.117	0.102	0.139	0.092	0.119	0.129	0.126
	Alpaca-7B	0.337	0.197	0.269	0.269	0.277	0.156	0.131	0.131	0.160	0.250	0.218
	Vicuna-7B	0.211	0.165	0.226	0.186	0.217	0.160	0.119	0.119	0.144	0.197	0.175
	Phoenix-7B	0.298	0.249	0.281	0.190	0.265	0.166	0.112	0.214	0.224	0.174	0.217
	ChatGPT	0.471	0.433	0.467	0.462	0.459	0.415	0.365	0.346	0.398	0.423	0.424
LLMs-FT (ours)	LLaMA-7B†	0.363	0.267	0.245	0.274	0.271	0.232	0.223	0.216	0.214	0.277	0.258
	LLaMA-2-7B†	**0.565**	0.484	0.510	0.506	0.523	0.436	0.416	0.355	0.378	0.478	0.465
	BLOOM-7B†	0.273	0.197	0.320	0.199	0.300	0.197	0.013	0.214	0.175	0.123	0.201
	Falcon-7B†	0.415	0.450	0.465	0.440	0.468	0.295	0.180	0.149	0.196	0.283	0.334
	Baichuan-2-7B†	0.541	**0.505**	0.515	0.501	0.513	0.453	0.444	0.388	0.412	0.480	0.475
	Alpaca-7B†	0.548	0.405	0.491	0.483	0.489	0.327	0.318	0.307	0.309	0.444	0.412
	Phoenix-7B†	0.481	0.435	0.461	0.366	0.465	0.323	0.264	0.410	0.435	0.334	0.397
Ensemble (ours)	LLaMA-7B + PoE†	0.476	0.443	0.448	0.462	0.466	0.431	0.423	0.371	0.425	0.442	0.439
	LLaMA-2-7B + PoE †	0.558	0.498	**0.518**	**0.520**	**0.528**	**0.470**	**0.455**	0.406	0.444	**0.494**	**0.489**
	BLOOM-7B + PoE†	0.485	0.444	0.461	0.460	0.474	0.425	0.418	0.376	0.431	0.440	0.441
	Falcon-7B + PoE†	0.494	0.479	0.485	0.488	0.499	0.419	0.400	0.355	0.411	0.437	0.447
	Baichuan-2-7B + PoE†	0.544	0.500	0.508	0.504	0.514	0.464	**0.455**	0.416	0.447	0.484	0.484
	Alpaca-7B + PoE†	0.543	0.461	0.503	0.504	0.511	0.420	0.412	0.387	0.413	0.476	0.463
	Phoenix-7B + PoE†	0.503	0.463	0.479	0.451	0.487	0.410	0.388	**0.420**	**0.455**	0.426	0.448

* Zhang, C., D'Haro, L. F., Zhang, Q., Friedrichs, T., & Li, H. (2023). Poe: A panel of experts for generalized automatic dialogue assessment. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *31*, 1234-1250.

Resultados a nivel de diálogo usando LLMs

- Conclusiones muy similares a turno.
- ChatGPT es muy bueno sin necesidad de ajuste fino. Llama2-7B mejora tras el ajuste fino.
- La combinación con nuestro modelo previo (FineD*) mejora los resultados.
- Las correlaciones aún siguen siendo muy bajas.

		Dialogue-Level										
BERT-Based	FineD†	0.386	0.354	0.362	0.362	0.372	0.346	0.341	0.343	0.339	0.376	0.358
LLMs-Zeroshot	LLaMA-7B	0.190	0.190	0.226	0.196	0.151	0.141	0.120	0.027	0.035	0.151	0.143
	LLaMA-2-7B	0.036	0.193	0.154	0.091	0.166	0.125	0.165	0.027	0.128	0.127	0.121
	BLOOM-7B	0.071	0.212	0.063	0.063	0.122	0.104	0.058	0.097	0.122	0.078	0.099
	Falcon-7B	0.286	0.240	0.248	0.268	0.153	0.113	0.107	0.134	0.168	0.219	0.194
	Baichuan-2-7B	0.296	0.316	0.270	0.258	0.274	0.211	0.198	0.156	0.201	0.235	0.241
	Alpaca-7B	0.441	0.321	0.386	0.404	0.402	0.301	0.268	0.208	0.270	0.356	0.336
	Vicuna-7B	0.347	0.234	0.243	0.260	0.242	0.209	0.220	0.132	0.148	0.231	0.226
	Phoenix-7B	0.312	0.292	0.264	0.261	0.291	0.254	0.163	0.253	0.253	0.206	0.255
	ChatGPT	0.419	0.375	0.407	0.395	0.404	0.378	0.310	0.324	**0.385**	0.363	0.376
LLMs-FT (ours)	LLaMA-7B†	0.237	0.201	0.192	0.208	0.240	0.173	0.169	0.151	0.172	0.207	0.195
	LLaMA-2-7B†	0.444	0.401	0.405	0.407	0.410	0.363	0.359	0.319	0.343	0.404	0.386
	BLOOM-7B†	0.289	0.235	0.269	0.249	0.253	0.175	0.132	0.288	0.274	0.136	0.230
	Falcon-7B†	0.376	0.366	0.314	0.334	0.320	0.231	0.146	0.142	0.197	0.174	0.260
	Baichuan-2-7B†	0.344	0.329	0.309	0.315	0.316	0.275	0.323	0.278	0.325	0.304	0.312
	Alpaca-7B†	0.420	0.362	0.383	0.394	0.379	0.309	0.263	0.255	0.278	0.351	0.339
	Phoenix-7B†	0.339	0.324	0.328	0.293	0.321	0.275	0.229	0.321	0.316	0.259	0.300
Ensemble (ours)	LLaMA-7B + FineD†	0.405	0.364	0.371	0.368	0.379	0.353	0.349	0.349	0.346	0.384	0.367
	LLaMA-2-7B + FineD †	**0.477**	**0.434**	**0.434**	**0.436**	**0.442**	**0.399**	**0.394**	**0.380**	**0.385**	**0.438**	**0.422**
	BLOOM-7B + FineD†	0.405	0.373	0.384	0.374	0.387	0.348	0.341	0.374	0.370	0.373	0.373
	Falcon-7B + FineD†	0.445	0.413	0.397	0.403	0.407	0.356	0.345	0.341	0.346	0.377	0.383
	Baichuan-2-7B + FineD†	0.402	0.379	0.366	0.371	0.374	0.339	0.369	0.333	0.369	0.364	0.367
	Alpaca-7B + FineD†	0.461	0.407	0.425	0.434	0.427	0.369	0.347	0.342	0.357	0.410	0.398
	Phoenix-7B + FineD†	0.403	0.373	0.377	0.356	0.379	0.340	0.317	0.368	0.363	0.338	0.361

* Chen Zhang, Luis Fernando D'Haro, Qiquan Zhang, Thomas Friedrichs, and Haizhou Li. 2022a. FineDeval: Fine-grained automatic dialogue-level evaluation. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 3336–3355, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Experimentos con LLMs: Multidominio y Multidimensional

Turn-Level Datasets	#Data	#Utt	Doc Len	IAA Range	Reused Annotations	Missing Annotations
Persona-USR (2020b)	300	9.3	98.4 / 12.0	0.3 ~ 0.7	rel, int, und, ovr	spe
Persona-Zhao (2020)	900	5.1	48.8 / 11.5	> 0.7	ovr	rel, int, und, spe
DailyDialog-Zhao (2020)	900	4.7	47.5 / 11.0	> 0.7	rel, ovr	int, und, spe
Topical-USR (2020b)	360	11.2	236.3 / 22.4	0.5 ~ 0.7	rel, int, und, ovr	spe
FED-Turn (2020a)	375	10.4	87.3 / 13.3	0.5 ~ 0.8	rel, int, spe, und, ovr	-
ConTurE-Turn (2022)	1066	3.8	21.7 / 11.0	~ 0.3	ovr	rel, int, und, spe
Dialogue-Level Datasets	**#Data**	**#Utt**	**Doc Len**	**IAA Range**	**Reused Annotations**	**Missing Annotations**
IEval-Dial (2022)	500	6.0	74.4	-	-	coh, eng, inf, div, ovr
Persona-See (2019)	500	12.0	91.2	-	-	coh, eng, inf, div, ovr
Reliable-Dial (2022)	500	21.2	178.1	-	-	coh, eng, inf, div, ovr
ConTurE-Dial (2022b)	119	17.9	153.9	-	-	coh, eng, inf, div, ovr
FED-Dial (2020a)	125	12.7	116.8	0.7 ~ 0.8	coh, eng, inf, div, ovr	-
Human-Eval (2022)	286	12.0	139.2	-	-	coh, eng, inf, div, ovr

- Prueba con 12 bb.dd anotadas por humanos. Dimensiones no incluidas originalmente “pseudo-anotadas” con GPT4 con 5 generaciones distintas (“acuerdo inter-anotador”: ~0.6) + correlación con evaluación humana usando la bb.dd FED: 0.85.

Zhang, C., D’Haro, L. F., Chen, Y., Zhang, M., & Li, H. (2024). A Comprehensive Analysis of the Effectiveness of Large Language Models as Automatic Dialogue Evaluators. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(17), 19515-19524. https://doi.org/10.1609/aaai.v38i17.29923

Resultados de LLMs (abiertos y cerrados) + Datos solo humanos

- Los modelos de OpenAI son los mejores tanto en datos pseudo-anotados (IZQ), como totalmente anotados con humanos (DER).
- Sin embargo, todos presentan correlaciones bajas.

	Turn-Level						Dialogue-Level					
Models	**Rel**	**Spe**	**Int**	**Und**	**Ovr**	**Avg**	**Coh**	**Eng**	**Div**	**Inf**	**Ovr**	**Avg**
LLaMA-7B	0.114	0.056	0.073	0.113	0.018	0.075	0.341	0.142	0.049	0.166	0.109	0.161
LLaMA-13B	0.442	0.240	0.333	0.348	0.402	0.353	0.522	0.425	0.237	0.324	0.404	0.382
LLaMA-2-7B	0.241	0.187	0.178	0.134	0.259	0.200	0.225	0.227	0.044	0.261	0.292	0.210
LLaMA-2-13B	0.303	0.122	0.161	0.269	0.225	0.216	0.318	0.362	-0.068	0.343	0.417	0.274
XGen-8K-7B	0.214	0.152	0.173	0.119	0.218	0.175	0.419	0.431	0.325	0.387	0.456	0.404
Falcon-7B	0.133	0.061	0.174	0.101	0.104	0.114	0.366	0.476	0.329	0.341	0.413	0.385
MPT-8K-7B	0.209	0.048	0.123	0.197	0.172	0.150	0.126	0.221	0.112	0.187	0.088	0.147
OpenLLaMA-7B	0.158	0.008	0.105	0.154	0.184	0.122	0.402	0.452	0.270	0.336	0.429	0.378
OpenLLaMA-13B	0.231	0.112	0.175	0.184	0.226	0.186	0.425	0.506	0.239	0.342	0.471	0.397
Pythia-7B	0.145	0.097	0.083	0.106	0.146	0.115	0.079	0.204	0.026	0.094	0.220	0.124
BLOOM-7B	0.082	0.021	0.124	0.099	0.122	0.089	0.224	0.271	0.137	0.217	0.277	0.225
LLaMA-2-Chat-7B	0.434	0.191	0.243	0.251	0.240	0.272	0.534	0.485	0.375	0.447	0.498	0.468
LLaMA-2-Chat-13B	0.559	0.353	0.207	0.320	0.380	0.364	0.644	0.610	0.228	0.514	0.613	0.522
Alpaca-7B	0.430	0.293	0.253	0.267	0.386	0.326	0.586	0.624	0.372	0.477	0.566	0.525
Vicuna-7B	0.368	0.096	0.221	0.100	0.219	0.201	0.490	0.468	0.279	0.488	0.482	0.441
Vicuna-13B	0.400	0.318	0.229	0.224	0.309	0.296	0.515	0.420	0.306	0.419	0.417	0.415
Falcon-Ins-7B	0.272	0.152	0.293	0.179	0.246	0.228	0.504	0.513	0.375	0.448	0.500	0.468
Tulu-13B	0.585	0.427	0.369	0.350	0.488	0.444	0.659	0.661	0.326	0.518	0.657	0.564
Chimera-7B	0.489	0.276	0.373	0.309	0.368	0.363	0.563	0.599	0.439	0.525	0.607	0.547
Chimera-13B	0.547	0.449	0.377	0.366	0.404	0.428	0.582	0.671	0.432	0.585	0.563	0.567
Phoenix-7B	0.314	0.101	0.291	0.258	0.234	0.240	0.480	0.493	0.146	0.334	0.416	0.374
Oasst-sft-Pythia-12B	0.144	0.028	0.203	0.132	0.110	0.123	0.386	0.358	0.236	0.346	0.423	0.350
Baize-v2-13B	0.477	0.350	0.333	0.353	0.337	0.370	0.568	0.544	0.397	0.482	0.469	0.492
Dolly-v2-12B	0.030	-0.009	0.061	-0.004	0.020	0.020	0.182	0.238	0.071	0.139	0.105	0.147
MPT-8K-7B-Instruct	0.139	0.092	0.176	0.095	0.099	0.120	0.321	0.352	0.316	0.308	0.315	0.322
XGen-8K-7B-Inst	0.272	0.293	0.267	0.145	0.265	0.248	0.502	0.515	0.308	0.417	0.506	0.450
ChatGLM-v2-6B	0.368	0.267	0.191	0.181	0.184	0.238	0.214	0.236	0.262	0.248	0.359	0.264
WizardLM-13B-V1.2	0.463	0.422	0.390	0.245	0.314	0.367	0.572	0.580	0.455	0.513	0.632	0.550
Palm-2 (text-bison-001)	0.666	0.563	0.454	0.422	0.601	0.541	0.649	0.674	0.473	0.557	0.674	0.605
ChatGPT (gpt-3.5-turbo)	0.595	0.578	0.518	0.536	0.542	0.554	0.724	0.705	0.516	0.568	0.707	0.644

Turn-Level						
	Rel	**Spe**	**Int**	**Und**	**Ovr**	**Avg**
Baize	0.449	0.147	0.302	0.290	0.337	0.305
Tulu	0.544	0.193	0.254	0.324	0.488	0.361
Chimera	0.507	0.234	0.312	0.316	0.404	0.354
Palm-2	0.616	0.317	0.343	0.384	0.601	0.452
ChatGPT	0.576	**0.408**	0.446	0.424	0.542	0.479
GPT-4	**0.704**	0.342	**0.538**	**0.558**	**0.677**	**0.564**
Dialogue-Level						
	Coh	**Eng**	**Div**	**Inf**	**Ovr**	**Avg**
Tulu	0.668	0.629	0.414	0.584	0.681	0.595
Chimera	0.595	0.628	0.507	0.609	0.525	0.573
WizardLM	0.536	0.522	0.477	0.540	0.605	0.536
Palm-2	0.584	0.633	0.550	0.604	0.614	0.597
ChatGPT	0.650	0.647	**0.551**	0.570	0.715	0.627
GPT-4	**0.760**	**0.689**	0.534	**0.620**	**0.744**	**0.669**

Recursos en Castellano

Base de datos de diálogo en arte

- Metodología de uso de LLMs para generación de diálogos sintéticos.
- Múltiples escenarios de diálogo:
  - 800 obras de arte, 378 artistas, 26 estilos de arte.
  - 13k diálogos.
  - 8 emociones extraídas desde ArtEmis.
- Datos en inglés y castellano
  - Disponibles en: https://huggingface.co/datasets/Astound/Art-GenEvalGPT y eCienciaDatos.
- Evaluación automática de Calidad: métricas objetivas y subjetivas + LLMs.
- Modelos: ChatGPT 3.5 y GPT4.

Conclusiones

La tarea de evaluación automática es difícil debido a la falta de datos, anotaciones y dificultad de tener información sobre diferentes dimensiones.

Se requieren nuevas bbdd y anotaciones en otras dimensiones: E.g., toxicidad, sesgo, coherencia, alucinación, sentido común, compromiso, cuestiones culturales y lingüísticas.

Datasets preparados por el grupo

DSTC10-T5: (Zhang et al., 2021): https://github.com/e0397123/dstc10_metric_track

DSTC11-T4 (Rodríguez-Cantelar et al., 2023): mario-rc/dstc11.t4 · Datasets at Hugging Face (incluye Español)

Art-GenEvalGPT: https://huggingface.co/datasets/Astound/Art-GenEvalGPT (incluye Español)

- Chen, Z., Sadoc, J., D'Haro, L. F., Banchs, R., & Rudnicky, A. (2021). Automatic evaluation and moderation of open-domain dialogue systems. *arXiv preprint arXiv:2111.02110*.
- Mario Rodríguez-Cantelar, Chen Zhang, Chengguang Tang, Ke Shi, Sarik Ghazarian, João Sedoc, Luis Fernando D'Haro, and Alexander I. Rudnicky. 2023. Overview of Robust and Multilingual Automatic Evaluation Metricsfor Open-Domain Dialogue Systems at DSTC 11 Track 4. In *Proceedings of The Eleventh Dialog System Technology Challenge*, pages 260–273, Prague, Czech Republic. Association for Computational Linguistics.