E2.1V1 MÉTODOS DE INYECCIÓN DE CONOCIMIENTO EN LLM

KG4LLM - SERVICIOS PARA EL ENRIQUECIMIENTO DE MODELOS DE LENGUAJE CON GRAFOS DE CONOCIMIENTO (SER-21/23 OTT)

Resumen

Este entregable consiste en la primera versión del software resultante de la implementación de los métodos desarrollados para la inyección de conocimiento en LLM en el marco de PT2. Esta iteración se centra en resultados preliminares en el desarrollo de métodos basados en Direct Preference Optimization (DPO) para el alineamiento de LLM con un dataset de preferencias de factualidad construido a partir de recursos de dominio, con el fin de aumentar el número de hechos factuales a la salida del modelo y reducir el número de hechos incorrectos

José Manuel Gómez Pérez
Cristian Berrío

30 de Junio de 2024
Expert.ai Language Technology Research Lab

Historia de revisions

Revision	Date	Description	Author (Organisation)
0.1	30/04/2024	Tabla de contenidos y estructura básica	Expert.ai
0.2	01/05/2024	Primera versión completa	Expert.ai
1.0	30/06/2024	Versión final	Expert.ai

Tabla de contenidos

1 Introducción

Este entregable recoge la primera iteración de los resultados del paquete de trabajo PT2 en el proyecto KG4LLM, sobre el desarrollo de métodos de inyección de conocimiento en LLM previamente definidos en la tarea T1.2 e introducidos en el entregable E1.1. Este entregable se basa en las guías definidas por E1.1 y tiene en cuenta los retos y limitaciones allí identificados, así como las recomendaciones basadas en dicho análisis.

El resto de este documento se estructura de la manera siguiente. La sección **Error! Reference source not found.** introduce de manera general la filosofía detrás del desarrollo de los métodos de inyección de conocimiento en LLM y las líneas generales que seguimos a la hora de presentarlo en este documento. La sección **Error! Reference source not found.** presenta el método de inyección de conocimiento propuesto, que se basa en el marco de evaluación presentado en E5.1 y el algoritmo de Direct Preference Optimization (DPO). Los datasets de preferencias de factualidad sintetizados en el marco de este entregable utilizados para ilustrar la aplicación de este método en el dominio de seguros se presentan en la sección 4, así como el método seguido para generar esos datasets. A continuación, la sección 5 describe el proceso seguido para entrenar un LLM objetivo sobre un dataset de preferencias de factualidad, siguiendo el algoritmo DPO. La sección 6 presenta el repositorio en el que software, modelos y datos se han puesto a disposición de la comunidad de usuarios de INESData. Finalmente, la sección 7 concluye este documento y esboza parte del trabajo futuro.

2 Enfoque

El análisis llevado a cabo en el entregable E4.1 puso de manifiesto la escasez de recursos abiertos disponibles para llevar a cabo los objetivos del proyecto KG4LLM. Estos recursos son necesarios para alimentar los métodos existentes de inyección de conocimiento estructurado en LLM con el propósito de su adaptación a dominio en el marco lingüístico del español y lenguas cooficiales. Métodos como K-Adapter (Wang et al., 2021) requieren de un grafo de conocimiento y un gran corpus de texto anotado en el que se haya etiquetado previamente las entidades y relaciones del grafo que aparecen en el corpus. Este es un proceso costoso en términos de tiempo y esfuerzo, que se complica aún más en un escenario vertical en el que escasean los grafos de conocimiento u otros recursos semánticos específicos de dominio, los corpora de texto y datasets de verificación y los conjuntos de datos de evaluación en tareas downstream NLP. Más aún en escenarios multilingües como el propuesto en el marco general de INESData. Por otro lado, K-Adapter y métodos similares fueron diseñados para arquitecturas Transformer basadas en encoders y no han sido adaptados a los modernos LLM, basados en una arquitectura decoder y con varios órdenes de magnitud más en cuanto a número de parámetros entrenables.

Por otro lado, recientes avances en el campo de la investigación en NLP para la mejora de la factualidad en LLM han causado un giro hacia nuevos enfoques que se alejan de los ya tradicionales métodos de inyección de conocimiento estructurado durante las sucesivas fases de preentrenamiento del LLM. Entre estos avances, cabe destacar de manera significativa el algoritmo DPO (Direct Preference Optimization) propuesto por Rafailov et al. (2023). DPO permite alinear el texto generado por el LLM con un dataset de preferencias relativas a dimensiones como seguridad, ausencia de lenguaje tóxico o efectividad de la respuesta generada, sin necesidad de entrenar un modelo de recompensa (Christiano et al., 2017). Estudios recientes como (Tian et al., 2023b) han demostrado que DPO es también un método efectivo para alinear LLM con un objetivo de factualidad como una preferencia más en el proceso de alineamiento del LLM.

Basados en este trabajo y con el fin de afrontar los retos de adaptación a dominio y multilingüismo en escenarios de bajos recursos, los métodos propuestos en este entregable tienen como objetivo aumentar la proporción de hechos factuales generados por un LLM mientras se reduce la cantidad de alucinaciones o hechos no factuales, utilizando métodos de alineamiento con preferencias mediante el algoritmo DPO. Estos métodos buscan ajustar los parámetros de un LLM sobre un objetivo de entrenamiento cuya función de pérdida se centra en optimizar la métrica de factualidad calculada mediante el marco de trabajo de evaluación presentado en el entregable E5.1. Este método aporta además flexibilidad con respecto de la disponibilidad de recursos estructurados para inyectar conocimiento factual de dominio en el LLM en los distintos idiomas objetivo. Dichos recursos pueden incluir, de menor a mayor grado de expresividad, listas de entidades, diccionarios, tesauros, taxonomías o grafos de conocimiento estructurados semánticamente.

3 Inyección de conocimiento en LLM mediante DPO

Una vez contamos con un método que nos permite estimar la factualidad de un LLM, definido por el marco de evaluación presentado en el entregable E5.1, es posible construir un conjunto de datos de preferencias para ajustar la factualidad de un LLM determinado a partir de un conjunto de prompts sin etiquetar. Utilizando los estimadores basados en referencias o en la confianza del modelo descritos en E5.1, y basándonos en el trabajo original de Tian et al. (2023b), utilizamos el método llamado FacTune, que incluye dos variantes: FacTune-FS, en el que se utiliza FactScore como estimador de factualidad basado en referencias, y FacTune-MC, basado en la confianza del modelo en la respuesta generada.

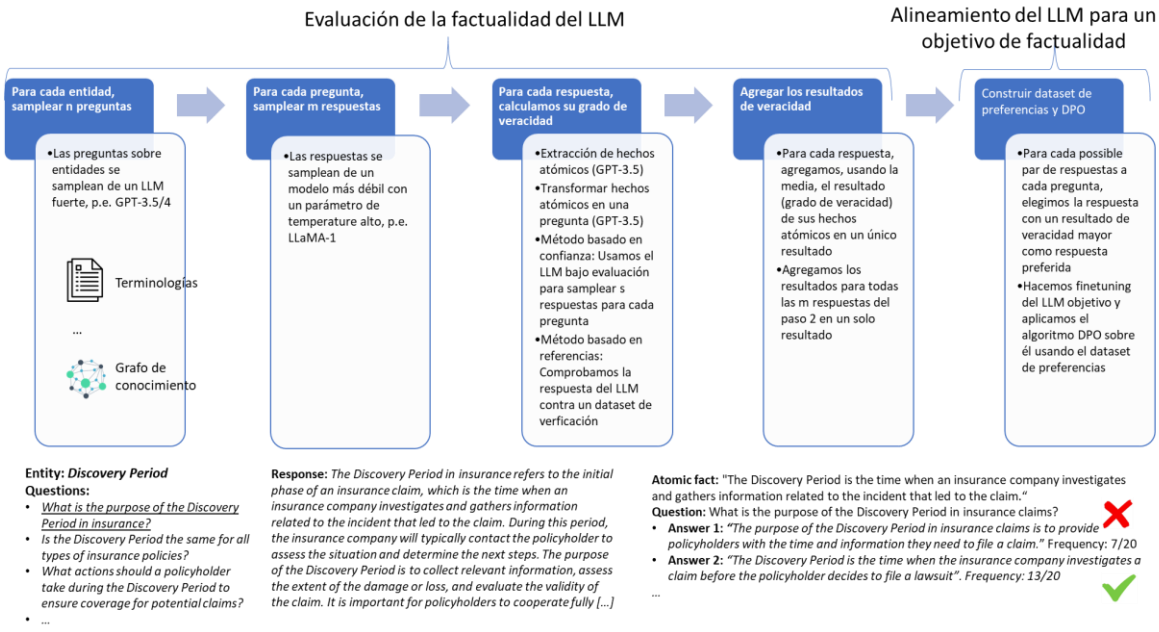

Figura 1: Pipeline completo de inyección de conocimiento factual en LLM procedente de un recurso de conocimiento externo. Las cuatro primeras etapas corresponden a la evaluación de factualidad del LLM, mientras que la última se centra en el desarrollo de un dataset de preferencias para alinear el modelo con un objetivo de factualidad y en el entrenamiento del modelo sobre ese dataset siguiendo el algoritmo DPO.

Para cada pregunta generada en el paso inicial descrito en la Figura 1, elegimos como respuesta preferida para el conjunto de datos de preferencias aquella que tiene una puntuación de factualidad más alta según el estimador que se esté utilizando y, como respuestas rechazadas,

todas las demás respuestas generadas por el LLM para esa pregunta. A continuación, construimos el dataset de preferencias, formado por pares *<respuesta preferida, respuesta rechazada>* asociados a cada pregunta. Finalmente, se entrena el LLM, ajustando sus parámetros mediante el pipeline de entrenamiento DPO aplicado sobre ese dataset, usando todas las respuestas del modelo como objetivo de entrenamiento en esta fase de ajuste supervisado o SFT (del inglés, supervised fine-tuning).

4 Datasets de DPO

Siguiendo el procedimiento explicado en la sección anterior generamos el dataset de preferencias para entrenar un LLM dado siguiendo el algoritmo de DPO. Inicialmente usamos 6 respuestas por prompt a la hora de samplear el modelo. Sin embargo, nos interesa saber cómo influye un mayor o menor valor en el resultado final del modelo entrenado con DPO, por lo que hemos generado varias versiones del dataset, utilizando diferente número de respuestas por prompt: 5, 7, 10, 15, 20 y 25.

4.1 Dataset basado en la confianza del modelo

A continuación, se presentan en la Tabla 1 las diferentes versiones del dataset de preferencias utilizando la confianza del modelo como estimación de la factualidad de las diferentes respuestas. Como se puede observar, el tamaño final varía considerablemente variando ligeramente el número de muestras por prompt. Se puede observar que el número de pares en el dataset de DPO tiende a duplicarse a cada paso, hasta llegar a las 20 muestras por prompt.

# muestras por prompt	# muestras	# pares en el dataset DPO
5	2730	5404
7	3822	11353
10	5460	24337
15	8190	56791
20	10920	102615
25	13650	162023

Tabla 1 Diferentes versiones del dataset de preferencias para el entrenamiento con DPO.

El número de muestras será el tamaño del dataset que vamos a tener para entrenar el LLM a través de un primer *supervised finetuning (SFT)*, para posteriormente entrenar con DPO.

4.2 Dataset basado en referencias.

En el caso de la generación del dataset para DPO basado en referencias, hemos inicialmente utilizado 7 muestras por prompt, pero esto se extenderá a los mismos números de muestras por prompt usados para generar el dataset de DPO basado en la confianza del modelo. En la Tabla 2 se incluyen los tamaños de las diferentes versiones.

# muestras por prompt	# pares en el dataset DPO
7	8167

Tabla 2 Diferentes versiones del dataset de preferencias para el entrenamiento con DPO basado en referencias.

Se puede ver que el tamaño del dataset en este caso es ligeramente inferior al número de pares basado en la confianza del modelo. Esto se debe a que hay cuatro entidades en el conjunto de entrenamientos que no han podido ser mapeadas a ningún artículo de Wikipedia, por lo que no se puede evaluar la factualidad para esos casos, en los que hay que descartar estas entidades para la generación del dataset de DPO.

5 Entrenamiento DPO

Tanto para el entrenamiento mediante *supervised finetuning* como DPO, se ha utilizado la librería TRL[1] (Transformer Reinforcement Learning) de HuggingFace. Esta librería cuenta con ejemplos[2] de código para entrenar un Llama-2 con el algoritmo de DPO, en los que nos hemos basado para entrenar utilizando los datasets generados.

Para el *supervised finetuning* se han reutilizado los parámetros que vienen en el código de ejemplo, aunque cambiando los max steps para ajustarse a los diferentes tamaños de dataset que manejamos. Para el caso del entrenamiento con DPO, se han aplicado los hiperparámetros propuestos por Rafailov et. al (2023): un tamaño de *batch* efectivo de 64, un *learning rate* de 1e-6, y *constant_with_warmup* para el tipo de *learning rate scheduler*. El resto de los parámetros se han mantenido, aunque ajustando los max_steps, para entrenar durante un máximo de 4 epochs. El checkpoint con el *validation loss* más bajo es el que se ha empleado para evaluar posteriormente su FactScore. A continuación, se presentan los resultados obtenidos.

Modelo	# muestras por prompt	método	# hechos correctos	# hechos incorrectos	FactScore
Llama-2	5	SFT	4.80	1.29	0.7999
		FactTune-MC	4.49	1.31	0.7836
	7	SFT	4.76	1.24	0.7932
		FactTune-MC	4.56	1.20	0.7981
		FactTune-FS	4.68	1.31	0.7862
	10	SFT	4.56	1.35	0.7840
		FactTune-MC	4.60	1.52	0.7760
	15	SFT	4.78	1.42	0.7840
		FactTune-MC	4.69	1.50	0.7809
	20	SFT	4.71	1.36	0.7828
		FactTune-MC	4.83	1.20	**0.8018**
	25	FactTune-MC	4.68	1.42	0.7788

[1] https://huggingface.co/docs/trl/en/index

[2] https://github.com/huggingface/trl/tree/main/examples/research_projects/stack_llama_2/scripts

Tabla 3 Resultados obtenidos al entrenar usando el algoritmo de DPO con las diferentes versiones del dataset de preferencias.

Observamos que utilizando 20 muestras por prompt se consigue mejorar la factualidad con respecto al modelo SFT. Así mismo se puede apreciar que existe una relación directa entre el número de muestras por prompt y el número de hechos correctos obtenido por FactTune-MC, excepto para 25 muestras por prompt, en el que este valor cae con respecto al valor de 20 muestras por prompt. En futuros experimentos, se extenderá el número de respuestas por pregunta a 30 y 40, para ver si hay alguna variación en las tendencias encontradas hasta el momento. A la hora de comparar los resultados obtenidos utilizando el dataset basado en referencias, FactTune-FS, con el basado en la confianza del modelo, Factune-MC, para el caso de 7 respuestas por prompt, se ve que el resultado de Factune-FS es peor. Es posible que este peor resultado sea debido a un tamaño de dataset más pequeño o una calidad o cobertura del dataset de verificación insuficiente. En futura experimentación se extenderá el entrenamiento basado en referencias con otros tamaños para ver si esta tendencia se mantiene.

6 Repositorio

El código utilizado para entrenar tanto el *supervised finetuning* como con el algoritmo DPO se encuentra en el repositorio https://github.com/oeg-upm/inesdata-knowledge-injection. Inicialmente está pensado para entrenar un Llama-2, aunque posteriormente se harán los ajustes necesarios para soportar otros modelos. Los pasos a seguir son los siguientes:

1. Hacer un *supervised finetuning* con el script “sft_llama2.py”.
2. Entrenar el algoritmo de DPO con el script “dpo_llama2.py”, partiendo del modelo SFT entrenado en el paso anterior.

Previamente es necesario haber generado el correspondiente dataset para el *supervised finetuning*, y el dataset de preferencias, ya sea basado en la confianza del modelo como basado en referencias. Para esto nos apoyamos en código del repositorio del marco de evaluación de la factualidad, que se encuentra disponible en https://github.com/oeg-upm/inesdata-fact-eval.

7 Conclusiones y trabajo futuro

El método de inyección de conocimiento presentados en este entregable se basa en el marco de evaluación de factualidad propuesto en E5.1 para generar un dataset de preferencias de factualidad con el que entrenar el LLM mediante el algoritmo DPO. Este método formula la tarea de inyección de conocimiento en LLM como un problema de alineamiento de preferencias, en el que, mediante DPO, entrenamos el LLM a *preferir* la generación de texto factual y centrado en el dominio (en este caso, evaluamos sobre el dominio de seguros) sobre texto alucinatorio o genérico. El trabajo futuro incluye reforzar la evaluación que presentamos en este documento con más experimentación, que se extenderá sobre otros LLM, dominios y lenguas. Los resultados de este trabajo, así como los de E5.1 servirán de base para el diseño del demostrador (E5.2), que se irá poblando con nuevas evaluaciones y modelos modificados siguiendo este método o evoluciones suyas, según avancemos nuestra investigación.

Referencias

Ahn, J., & Oh, A. (2021). **Mitigating language-dependent ethnic bias in BERT**. arXiv preprint arXiv:2109.05704.

Almazrouei E., Alobeidli H., Alshamsi A., Cappelli A., Cojocaru R., Debbah M., Goffinet É., Hesslow D., Launay J., Malartic Q., Mazzotta D., Noune B., Pannier B., & Penedo G. (2023). **The Falcon Series of Open Language Models.** *arXiv preprint arXiv:2311.16867*

Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2023). **Self-rag: Learning to retrieve, generate, and critique through self-reflection**. arXiv preprint arXiv:2310.11511.

Bender, EM. (2019). **A typology of ethical risks in language technology with an eye towards where transparent documentation can help.**

Benjamins, V. R. (1993). **Problem Solving Methods for Diagnosis.** PhD thesis, University of Amsterdam, Amsterdam, The Netherlands.

Benjamins, V. R., & Fensel, D. (1998). **Problem-solving methods**. International Journal of Human-Computer Studies, 49(4), 305-313.

Berquand, A., Ladeira A. V. (2022). **From Mission Description to Knowledge Graph: Applying Transformer-based models to map knowledge from publicly available satellite datasets**.

Bommasani, R., Liang, P., & Lee, T. (2023**). Holistic Evaluation of Language Models. Annals of the New York Academy of Sciences**.

Bordia, S., & Bowman, S. R. (2019). Identifying and reducing gender bias in word-level language models. arXiv preprint arXiv:1904.03035.

Brown, TB. et al. (2020). **Language Models are Few-Shot Learners.** ArXiv, abs/2005.14165.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). **Semantics derived automatically from language corpora contain human-like biases**. *Science*, *356*(6334), 183-186.

Chen, J., Sriram, A., Choi, E., & Durrett, G. (2022). **Generating Literal and Implied Subquestions to Fact-check Complex Claims**. *In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 3495–3516). Association for Computational Linguistics.*

Chen S., Zhao Y., Zhang J., Chern I., Gao S., Liu P., & He J. (2023). **FELM: Benchmarking Factuality Evaluation of Large Language Models.** *In Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Track on Datasets and Benchmarks*

Chern, I. et al. (2023). **Factool: Factuality detection in generative ai – a tool augmented framework for multi-task and multi-domain scenarios.** CoRR, abs/2307.13528.

Chowdhery, A.- et al. (2022). **Palm: Scaling language modeling with pathways.** URL: https://arxiv.org/abs/2204.02311

Christiano, P.F., Leike, J., Brown, T.B., Martic, M., Legg, S., & Amodei, D. (2017). **Deep Reinforcement Learning from Human Preferences.** ArXiv, abs/1706.03741.

Chuang, Y., Xie, Y., Luo, H., Kim, Y., Glass, J.R., & He, P. (2023). **DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models.** ArXiv, abs/2309.03883.

Cobbe K., Kosaraju V., Bavarian M., Chen M., Jun H., Kaiser L., Plappert M., Tworek J., Hilton J., Nakano R., Hesse C., & Schulman J. (2021). **Training Verifiers to Solve Math Word Problems.** *arXiv preprint arXiv:2110.14168*

Costa-Jussa, Marta R. et al. (2022). **No Language Left Behind: Scaling Human-Centered Machine Translation**. ArXiv, abs/2207.04672

Dai, D. et al. (2022). **Knowledge Neurons in Pretrained Transformers.** In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Davison, J., Feldman, J., and Rush, A. (2019). **Commonsense knowledge mining from pretrained models.** In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 1173–1178, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1109

Delobelle, P., Tokpo, E. K., Calders, T., & Berendt, B. (2022). **Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models.** In NAACL 2022: the 2022 Conference of the North American chapter of the Association for Computational Linguistics: human language technologies (pp. 1693-1706).

Denaux, R., Gomez-Perez, JM. (2020). **Linked Credibility Reviews for Explainable Misinformation Detection.** In: J. Z. Pan, V. Tamma, C. d'Amato, K. Janowicz, B. Fu, A. Polleres, O. Seneviratne, L. Kagal (Eds.), The Semantic Web – ISWC 2020, Springer International Publishing, Cham, 2020, pp. 147–163.

Denaux, R. and Gomez-Perez, JM. (2019). **Vecsigrafo: Corpus-based Word-Concept Embeddings. Bridging the Statistic-Symbolic Representational Gap in Natural Language Processing.** Semantic Web Journal 10, 5 (2019), 881–908. https://doi.org/10.3233/SW-190361

Devlin, J. et al. 2019. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.** In Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol.1 (Long and Short Papers), pp. 4171-4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dhamala, J., Sun, T., Kumar, V., Krishna, S., Pruksachatkun, Y., Chang, K. W., & Gupta, R. (2021, March). **Bold: Dataset and metrics for measuring biases in open-ended language generation**. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 862-872).

Dhingra, B., Cole, JR., Eisenschlos, JM., Gillick, D., Eisenstein, J., and Cohen, WW. (2022). **Time-aware language models as temporal knowledge bases.** Transactions of the Association for Computational Linguistics, 10:257–273.

Dinan, E., Fan, A., Wu, L., Weston, J., Kiela, D., and Williams, A. (2020). Multidimensional gender bias classification. ArXiv.

Dolci, T., Azzalini, F., & Tanelli, M. (2023). **Improving Gender-Related Fairness in Sentence Encoders: A Semantics-Based Approach. Data Science and Engineering**, 1-19.

Driess, D. et al. (2023). **PaLM-E: An Embodied Multimodal Language Model.** International Conference on Machine Learning.

Fensel, D.A. (2000). **Problem-Solving Methods: Understanding, Description, Development, and Reuse.** Lecture Notes in Computer Science 1791, Springer 2000, ISBN 3-540-67816-6

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., ... & Ahmed, N. K. (2023). **Bias and fairness in large language models: A survey**. *arXiv preprint arXiv:2309.00770*.

García-Silva, A., Berrio, C., Gómez-Pérez, JM. (2023). **Textual Entailment for Effective Triple Validation in Object Prediction.** The Semantic Web – ISWC 2023, Springer International Publishing, Cham, 2020, to appear.

Gao L., Biderman S., Black S., Golding L., Hoppe T., Foster C., Phang J., He H., Thite A., Nabeshima N., Presser S., & Leahy C. (2020). **The Pile: An 800GB Dataset of Diverse Text for Language Modeling.** *arXiv preprint arXiv:2101.00027.*

Gao, L., Madaan, A., Zhou, S., Alon, U., Liu, P., Yang, Y., Callan, J., & Neubig, G. (2022). **PAL: Program-aided Language Models.** ArXiv, abs/2211.10435.

Gao, L., et al. (2023). **RARR: Researching and Revising What Language Models Say, Using Language Models.** In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

Gao, S., et al. (2024). **Efficient Tool Use with Chain-of-Abstraction Reasoning**. arXiv:2401.17464

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2023). **Retrieval-augmented generation for large language models: A survey**. arXiv preprint arXiv:2312.10997.

Geva M., Khashabi D., Segal E., Khot T., Roth D., & Berant J. (2021**). Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies.** *arXiv preprint arXiv:2101.02235*

Gómez-Pérez, JM., Ortega, R. (2023). **E4.1 Análisis y Definición de Dominios de Aplicación y Casos de Uso.** KG4LLM Technical Report.

Gómez-Pérez, JM., García-Silva, A., Leone, R., Albani, M., Fontaine, M., Poncet, C., Summerer, L., Donati, A., Roma, I., Scaglioni, S. (2023). **Artificial Intelligence and Natural Language Processing and Understanding in Space: A Methodological Framework and Four ESA Case Studies.** Engineering Applications of Artificial Intelligence (to appear).

Gomez-Perez, JM., Denaux, R., Garcia-Silva, A. (2020) **A Practical Guide to Hybrid Natural Language Processing - Combining Neural Models and Knowledge Graphs for NLP.** Springer, Cham. DOI: https://doi.org/10.1007/978-3-030-44830-1

Gómez-Pérez, JM., Ortega, R. (2020). **ISAAQ - Mastering Textbook Questions with Pre-trained Transformers and Bottom-Up and Top-Down Attention.** 5469-5479. 10.18653/v1/2020.emnlp-main.441. Empirical Methods to Natural Language Processing (EMNLP) 2020.

Gomez-Perez, Jose Manuel. (2010). **Acquisition and understanding of process knowledge using problem solving methods.** Studies on the Semantic Web, IOS Press, 978-1-60750-600-3 (print) | 978-1-61499-341-4 (online). DOI: https://doi.org/10.3233/978-1-61499-341-4-i

Guo, W., & Caliskan, A. (2021, July). **Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases**. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (pp. 122-133).

He, B. et al. (2020). **BERT-MK: Integrating graph contextualized knowledge into pre-trained language models.** In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2281–2290, Online, Nov. 2020. Association for Computational Linguistics.

He, X., Tian, Y., Sun, Y., Chawla, N., Laurent, T., LeCun, Y., Bresson, X., & Hooi, B. (2024). **G-Retriever: Retrieval-Augmented Generation for Textual Graph Understanding and Question Answering.** ArXiv, abs/2402.07630.

Hendrycks D., Burns C., Basart S., Zou A., Mazeika M., Song D., & Steinhardt J. (2021). **Measuring Massive Multitask Language Understanding.** *arXiv preprint arXiv:2009.03300.*

Hoffmann J. et al. (2022). **Training Compute-Optimal Large Language Models.** *arXiv preprint arXiv:2203.15556*

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., Laroussilhe, Q.D., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). **Parameter-Efficient Transfer Learning for NLP. International Conference on Machine Learning.**

Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., Yu, J.A., Joulin, A., Riedel, S., & Grave, E. (2022). **Few-shot Learning with Retrieval Augmented Language Models.** ArXiv, abs/2208.03299.

Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., & Grave, E. (2021). **Unsupervised dense information retrieval with contrastive learning**. arXiv preprint arXiv:2112.09118.

Ji, Z. et al. 2022. **Survey of hallucination in natural language generation.** ACM Computing Surveys.

Ji, H., Grishman, R.: **Knowledge Base Population: Successful Approaches and Challenges**. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 1148–1158. Association for Computational Linguistics, Portland, Oregon, USA (Jun 2011), https://aclanthology.org/P11-1115

Kadavath, S., Conerly, T., Askell, A., Henighan, T.J., Drain, D., Perez, E., Schiefer, N., Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., Ganguli, D., Hernandez, D., Jacobson, J., Kernion, J., Kravec, S., Lovitt, L., Ndousse, K., Olsson, C., Ringer, S., Amodei, D., Brown, T.B., Clark, J., Joseph, N., Mann, B., McCandlish, S., Olah, C., & Kaplan, J. (2022). **Language Models (Mostly) Know What They Know.** ArXiv, abs/2207.05221.

Kamoi, R., Goyal, T., Rodriguez, J.D., & Durrett, G. (2023). **WiCE: Real-World Entailment for Claims in Wikipedia.** ArXiv, abs/2303.01432.

Kaneko, M., & Bollegala, D. (2022, June). **Unmasking the mask–evaluating social biases in masked language models**. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 11, pp. 11954-11962).

Kandpal, N., Deng, H., Roberts, A., Wallace, E., Raffel, C. (2022). **Large Language Models Struggle to Learn Long-Tail Knowledge.**

Kirkpatrick, J. et al. 2017. **Overcoming catastrophic forgetting in neural networks.** Proceedings of the national academy of sciences, 114(13):3521–3526.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). **Large language models are zero-shot reasoners**. *Advances in neural information processing systems*, *35*, 22199-22213.

Komeili, M., Shuster, K., and Weston, J. (2022). **Internet-augmented dialogue generation.** In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8460–8478, Dublin, Ireland. Association for Computational Linguistics.

Kryscinski, W., McCann, B., Xiong, C., and Socher, R. (2020). **Evaluating the Factual Consistency of Abstractive Text Summarization.** In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 9332–9346, Online. Association for Computational Linguistics.

Kuhn, L., Gal, Y. and Farquhar, S. (2023) **Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation.**

Kurita, K., Vyas, N., Pareek, A., Black, A. W., & Tsvetkov, Y. (2019). **Measuring bias in contextualized word representations**. arXiv preprint arXiv:1906.07337.

Laurençon, H. et al. (2022). **The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset.** *In Advances in Neural Information Processing Systems (pp. 31809–31826). Curran Associates, Inc.*

Lauscher, A., Majewska, O., Ribeiro, L., Gurevych, I., Rozanov, N., Glavaš, G. (2020). **Common Sense or World Knowledge? Investigating Adapter-Based Knowledge Injection into Pretrained Transformers**. 43-49. 10.18653/v1/2020.deelio-1.5.

Lawrence, Peter. (2024). **Text-to-Graph via LLM: pre-training, prompting, or tuning?.** https://medium.com/@peter.lawrence_47665/text-to-graph-via-llm-pre-training-prompting-or-tuning-3233d1165360

Lawrence, Peter. (2023). **Large Language Model = Knowledge Graph Store? Yes, by Fine-Tuning LLM With KG**. https://betterprogramming.pub/large-language-model-knowledge-graph-store-yes-by-fine-tuning-llm-with-kg-f88b556959e6

Lehmann, J. et al. (2015). **DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia.** Semantic Web, 6, 167-195.

Levine, Y. et al. (2020). **SenseBERT: Driving some sense into BERT.** In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, July 2020.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). **Retrieval-augmented generation for knowledge-intensive nlp tasks**. Advances in Neural Information Processing Systems, 33, 9459-9474.

Li, K., Patel, O., Vi'egas, F., Pfister, H., & Wattenberg, M. (2023). **Inference-Time Intervention: Eliciting Truthful Answers from a Language Model.** ArXiv, abs/2306.03341.

Lin et al. (2022). **Few-shot Learning with Multilingual Generative Language Models.** In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lin, S., Hilton, J., and Evans, O. 2022. **TruthfulQA: Measuring How Models Mimic Human Falsehoods.** In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Liu, Y. et al. (2019). **RoBERTa: A Robustly Optimized BERT Pretraining Approach.** ArXiv, abs/1907.11692.

Liu, Y., Fabbri, A., Liu, P., Zhao, Y., Nan, L., Han, R., Han, S., Joty, S., Wu, C.S., Xiong, C., & Radev, D. (2023). **Revisiting the Gold Standard: Grounding Summarization Evaluation with Robust Human Evaluation.** *In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 4140–4170). Association for Computational Linguistics.*

Logan IV, R. L., Liu, N. F., Peters, M. E., Gardner, M., & Singh, S. (2019). ***Barack's wife hillary: Using knowledge-graphs for fact-aware language modeling****. arXiv preprint arXiv:1906.07241.*

Malaviya C, Lee S., Chen S., Sieber E., Yatskar M., & Roth D. (2023). **ExpertQA: Expert-Curated Questions and Attributed Answers.** *arXiv preprint arXiv:2309.07852.*

Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., and Hajishirzi, H. 2023. **When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories.** In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Manakul, P., Liusie, A. and Gales, M. 2023. **SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models.** In Proceedings of the 2023 Conference on Empirical

Methods in Natural Language Processing, pages 9004–9017, Singapore. Association for Computational Linguistics.

May, C., Wang, A., Bordia, S., Bowman, S. R., & Rudinger, R. (2019). **On measuring social biases in sentence encoders**. arXiv preprint arXiv:1903.10561.

Maynez, J., Narayan, S., Bohnet, B, and McDonald, R. (2020). **On Faithfulness and Factuality in Abstractive Summarization.** In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1906–1919, Online. Association for Computational Linguistics.

Mcdermott, J., 1988. **Preliminary Steps Toward a Taxonomy of Problem-Solving Methods.** Springer US, Boston, MA. pp. 225–256. DOI: https://doi.org/10.1007/978-1-4684-7122-9_8

Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). **Locating and Editing Factual Associations in GPT.** *In Advances in Neural Information Processing Systems (pp. 17359–17372). Curran Associates, Inc.*

Mesquita, F., Cannaviccio, M, Schmidek, J., Mirza, P., and Barbosa, D. (2019). **KnowledgeNet: A Benchmark Dataset for Knowledge Base Population.** In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 749–758, Hong Kong, China. Association for Computational Linguistics.

Miller, GA. (1994). **WordNet: A Lexical Database for English.** In Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994.

Min S., Shi W., Lewis M., Chen X., Yih W., Hajishirzi H., & Zettlemoyer L. (2023). **Nonparametric Masked Language Modeling.** *arXiv preprint arXiv:2212.01349*

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.T., Koh, P., Iyyer, M., Zettlemoyer, L., and Hajishirzi, H. (2023). **Factscore: Fine-grained atomic evaluation of factual precision in long form text generation**.

Muhlgay, D., Ram, O., Magar, I., Levine, Y., Ratner, N., Belinkov, Y., Abend, O., Leyton-Brown, K., Shashua, A., & Shoham, Y. (2023). **Generating Benchmarks for Factuality Evaluation of Language Models.** ArXiv, abs/2307.06908.

Nadeem, M., Bethke, A., & Reddy, S. (2020). StereoSet: **Measuring stereotypical bias in pretrained language models**. arXiv preprint arXiv:2004.09456.

Nadeem, M., Bethke, A., and Reddy, S. (2021). **StereoSet: Measuring stereotypical bias in pretrained language models.** In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 5356–5371, Online. Association for Computational Linguistics.

Nangia, N., Vania, C., Bhalerao, R., and Bowman, S. (2020). **CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models.** In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1953–1967, Online. Association for Computational Linguistics.

Névéol, A., Dupont, Y., Bezançon, J., & Fort, K. (2022, May). **French CrowS-Pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English**. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 8521-8531).

A. Newell, J. C. Shaw, and H. A. Simon. **Report on a general problem solving program**. In IFIP congress, volume 256, page 64. Pittsburgh, PA, 1959.

A. Newell, H. A. Simon, et al. **Human problem solving**. Prentice-Hall, 1972.

Ni J., Qu C., Lu J., Dai Z., Ábrego GH., Ma J., Zhao VY., Luan Y., Hall KB., Chang M., & Yang Y. (2021). **Large Dual Encoders Are Generalizable Retrievers**. *arXiv preprint arXiv:2112.07899*

Nozza, D., Bianchi, F., & Hovy, D. (2021). HONEST: **Measuring hurtful sentence completion in language models**. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics.

O'Hern, M.S. & Rindfleisch, A. (2010). **Customer Co-Creation: A Typology and Research Agenda.** In: Review of Marketing Research, vol. 6, p. 84-106.

Opitz, J. (2019). **Argumentative relation classification as plausibility ranking.** In Preliminary Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers,

pages 193–202, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

Ouyang L., Wu J., Jiang X., Almeida D., Wainwright CL, Mishkin P., Zhang C., Agarwal S., Slama K., Ray A., Schulman J., Hilton J., Kelton F., Miller L., Simens M., Askell A., Welinder P., Christiano P., Leike J., & Lowe R. (2022). **Training language models to follow instructions with human feedback.** *arXiv preprint arXiv:2203.02155*.

Parisi, A. & Zhao, Y. & Fiedel, N. (2022). **TALM: Tool Augmented Language Models.** 10.48550/arXiv.2205.12255.

Patel, A., Bhattamishra, S., and Goyal, N. (2021). **Are NLP models really able to solve simple math word problems?** In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2080–2094, Online. Association for Computational Linguistics.

Penedo, G. et al. (2023). **The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only.** ArXiv, abs/2306.01116.

Peters, ME., Neumann, M., Logan, R, Schwartz, R., Joshi, V, Singh, S., and Smith, N. (2019). **Knowledge Enhanced Contextual Word Representations.** In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 43–54, Hong Kong, China. Association for Computational Linguistics.

Petroni, F. et al. (2019). **Language Models as Knowledge Bases.** In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Puri, R., & Catanzaro, B. (2019). **Zero-shot Text Classification With Generative Language Models.** ArXiv, abs/1912.10165.

Qian, R., Ross, C., Fernandes, J., Smith, E., Kiela, D., & Williams, A. (2022). **Perturbation augmentation for fairer nlp**. *arXiv preprint arXiv:2205.12586*.

Qiao, S., Ou, Y., Zhang, N., Chen, X., Yao, Y., Deng, S., ... & Chen, H. (2022). **Reasoning with language model prompting: A survey**. arXiv preprint arXiv:2212.09597.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). **Language Models are Unsupervised Multitask Learners.**

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C.D., & Finn, C. (2023). **Direct Preference Optimization: Your Language Model is Secretly a Reward Model.** ArXiv, abs/2305.18290.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). **Squad: 100,000+ questions for machine comprehension of text**. arXiv preprint arXiv:1606.05250.

Rehm, G., & Way, A. (2023). **Strategic Research, Innovation and Implementation Agenda for Digital Language Equality in Europe by 2030.** European Language Equality. Springer Cham. https://doi.org/10.1007/978-3-031-28819-7_45

Rehm G. et al. (2023). **European Language Grid A Language Technology Platform for Multilingual Europe.** Springer Cham. https://doi.org/10.1007/978-3-031-17258-8

Sanh, Victor & Webson, Albert & Raffel, Colin & Bach, Stephen & Sutawika, Lintang & Alyafeai, Zaid & Chaffin, Antoine & Stiegler, Arnaud & Scao, Teven & Raja, Arun & Dey, Manan & Bari, M & Xu, Canwen & Thakker, Urmish & Sharma, Shanya & Szczechla, Eliza & Kim, Taewoon & Chhablani, Gunjan & Nayak, Nihal & Rush, Alexander. (2021). **Multitask Prompted Training Enables Zero-Shot Task Generalization.**

Sap, M. et al. (2019). **ATOMIC: an atlas of machine commonsense for if-then reasoning.** In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI'19/IAAI'19/EAAI'19). AAAI Press, Article 372, 3027–3035. https://doi.org/10.1609/aaai.v33i01.33013027

Santhanam K. et al. (2022). **ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction**. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3715–3734, Seattle, United States. Association for Computational Linguistics.

Scao, TL. et al. (2022). **BLOOM: A 176B-Parameter Open-Access Multilingual Language Model.** ArXiv, abs/2211.05100.

Schick, T., et al. (2023). **Toolformer: Language Models Can Teach Themselves to Use Tools.** ArXiv, abs/2302.04761.

Schick, T., Udupa, S., and Schutze, H. (2021). **Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP.** Transactions of the Association for Computational Linguistics. https://doi.org/10.1162/tacl_a_00434

Schick, T., Schutze, H. (2021). **Generating datasets with pretrained language models.** In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.555

Schreiber, G. et al. (1994). **CommonKADS: A comprehensive methodology for KBS development.** IEEE Expert. 9. 28-37. 10.1109/64.363263.

Schreiber, G. (2000). **Knowledge engineering and management: the CommonKADS methodology**. MIT press.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). **Proximal policy optimization algorithms**. arXiv preprint arXiv:1707.06347.

Shuster, K., Poff, S., Chen, M., Kiela, D., and Weston, J. (2021). **Retrieval Augmentation Reduces Hallucination in Conversation.** In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Speer, R., Chin, J., and Havasi, C. (2017). **ConceptNet 5.5: an open multilingual graph of general knowledge.** In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI'17). AAAI Press, 4444–4451.

Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., and Wang, H. (2020). **ERNIE 2.0: A continual pre-training framework for language understanding.** In The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pages 8968–8975. AAAI Press, 2020.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2018). **FEVER: a Large-scale Dataset for Fact Extraction and VERification.** In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Thoppilan, R. et al. (2022). **LaMDA: Language Models for Dialog Applications.** ArXiv, abs/2201.08239.

Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., & Manning, C.D. (2023). **Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback.** ArXiv, abs/2305.14975.

Tian, K., Mitchell, E., Yao, H., Manning, C.D., & Finn, C. (2023). **Fine-tuning Language Models for Factuality**. ArXiv, abs/2311.08401.

Touileb, S., Øvrelid, L., & Velldal, E. (2022, July). **Occupational biases in Norwegian and multilingual language models**. In Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP) (pp. 200-211).

Touvron, H. et al. (2023). **Llama 2: Open Foundation and Fine-Tuned Chat Models.** ArXiv, abs/2307.09288.

Vashishtha, A., Ahuja, K., & Sitaram, S. (2023). **On evaluating and mitigating gender biases in multilingual settings**. *arXiv preprint arXiv:2307.01503*.

Vaswani, A. et al. (2017). **Attention is All you Need.** NIPS.

Wang, R. et al. (2021). **K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters.** In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 1405–1418, Online. Association for Computational Linguistics.

Wang, A., Cho, K., and Lewis, M. (2020). **Asking and Answering Questions to Evaluate the Factual Consistency of Summaries.** In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5008–5020, Online. Association for Computational Linguistics.

Wang, L., Xu, W., Lan, Y., Hu, Z., Lan, Y., Lee, R.K., & Lim, E. (2023). **Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models.** Annual Meeting of the Association for Computational Linguistics.

Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Naik, A., Ashok, A., Dhanasekaran, A., Arunkumar, A., Stap, D., Pathak, E., Karamanolakis, G., Lai, H., Purohit, I., Mondal, I., Anderson, J., Kuznia, K., Doshi, K., Pal, K., Patel, M., Moradshahi, M., Parmar, M., Purohit, M., Varshney, N., Kaza, P., Verma, P., Puri, R., Karia, R., Doshi, S., Sampat, S., Mishra, S., Reddy A, S., Patro, S., Dixit, T., & Shen, X. (2022). **Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks**. *In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 5085–5109). Association for Computational Linguistics.*

Wang Y., Kordi Y., Mishra S., Liu A., Smith NA., Khashabi D., & Hajishirzi H. (2023). **Self-Instruct: Aligning Language Models with Self-Generated Instructions.** *arXiv preprint arXiv:2212.10560.*

Wang, L., Xu, W., Lan, Y., Hu, Z., Lan, Y., Lee, R. K. W., & Lim, E. P. (2023). **Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models**. arXiv preprint arXiv:2305.04091.

Webster, K., Wang, X., Tenney, I., Beutel, A., Pitler, E., Pavlick, E., ... & Petrov, S. (2020). **Measuring and reducing gendered correlations in pre-trained models**. arXiv preprint arXiv:2010.06032.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). **Chain-of-thought prompting elicits reasoning in large language models**. Advances in Neural Information Processing Systems, 35, 24824-24837

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022). **Emergent abilities of large language models**. arXiv preprint arXiv:2206.07682.

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. **The FAIR Guiding Principles for scientific data management and stewardship.** Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

Xiong, W., Du, J, Wang, W., and Stoyanov, V. (2020). **Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model.** In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020.

Yao, Y., Huang, S., Zhang, N., Dong, L., Wei, F., & Chen, H. (2022). **Kformer: Knowledge Injection in Transformer Feed-Forward Layers.** ArXiv, abs/2201.05742.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). **Tree of thoughts: Deliberate problem solving with large language models**, may 2023. *arXiv preprint arXiv:2305.10601*, *14*.

Hongbin Ye, Ningyu Zhang, Hui Chen, and Huajun Chen. 2022. **Generative Knowledge Graph Construction: A Review**. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 1–17, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yu, D. et al. (2022). **KG-FiD: Infusing Knowledge Graph in Fusion-in-Decoder for Open-Domain Question Answering.** In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 4961–4974.

Zamani, H., Diaz, F., Dehghani, M., Metzler, D., and Bendersky, M. (2022). **Retrieval-Enhanced Machine Learning.** In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22). Association for Computing Machinery, New York, NY, USA, 2875–2886. https://doi.org/10.1145/3477495.3531722

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M, and Liu, Q. (2019). **ERNIE: Enhanced language representation with informative entities.** In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1441–1451, Florence, Italy, July 2019. Association for Computational Linguistics.

Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., ... & Sun, L. (2023). **A comprehensive survey on pretrained foundation models: A history from bert to chatgpt**. *arXiv preprint arXiv:2302.09419*.