TruthfulQA: Measuring How Language Models Mimic Human Falsehoods - A Briefing Doc
https://paperswithcode.com/dataset/truthfulqa
Source: Stephanie Lin et al. (2022)
TruthfulQA is a benchmark to measure whether a language model is truthful in generating answers to questions. The authors crafted questions that some humans would answer falsely due to a false belief or misconception.

Core Problem: While large language models (LLMs) exhibit impressive fluency, they often generate false statements. This raises concerns about accidental and malicious misuse, and potentially hindering the adoption of LLMs in high-risk applications demanding factual accuracy.

Imitative Falsehoods: False statements generated by LLMs due to their tendency to mimic common misconceptions and falsehoods prevalent in their training data.
“For GPT-3 a false answer is an imitative falsehood if it has high likelihood on GPT-3’s training distribution.”

TruthfulQA Benchmark: A dataset of 817 questions designed to provoke imitative falsehoods. It spans 38 categories, including health, law, and politics.

From the Paper:

- **LLMs Struggle with Truthfulness:** Even advanced models like GPT-3 achieved only 58% accuracy on TruthfulQA, compared to 94% accuracy for humans.
  "The best model was truthful on 58% of questions, while human performance was 94%."
- **Larger Models, Less Truthful**: Larger models performed worse on TruthfulQA ("inverse scaling"). This challenges the assumption that scaling models improves performance.
  "In contrast to other NLP tasks, larger models are less truthful on TruthfulQA."
- **Informative but False Answers**: One of the most alarming findings is that models often generate answers that are both false and highly informative. These responses mimic common misconceptions and makes them deceptively convincing.
  "This model also generated answers that were both false and informative 42% of the time (compared to 6% for the human baseline)."
- **Automated Metric (GPT-judge)**: Fine-tuned GPT-3-based model to evaluate the truthfulness of answers. It achieved 90-96% accuracy in predicting human evaluations. It is a scalable alternative to costly human reviews.
  "Automated metric predicts human evaluation with high accuracy."

Interpreting the Results: The adversarial nature of TruthfulQA questions likely contributes to the poor performance of LLMs. The inverse scaling trend underscores that simply increasing model size isn't enough to improve truthfulness.

Mitigating Imitative Falsehoods: Achieving truthful LLMs requires moving beyond just scaling models. Potential strategies include:

- **Prompt Engineering**: Crafting prompts that explicitly instruct models to prioritize truthfulness.
- **Fine-Tuning with Targeted Data**: Using datasets specifically designed to promote truthfulness or leveraging human feedback via reinforcement learning.
- **Information Retrieval**: Integrating reliable information retrieval mechanisms to ensure accurate responses.

On Faithfulness and Factuality in Abstractive Summarization
https://paperswithcode.com/paper/on-faithfulness-and-factuality-in-abstractive
Source: Maynez et al. (2020)
The paper focuses on the problem of **hallucinations** in neural abstractive summarization models the generation of content that is not supported by the input document.

From the Paper:

- **Hallucinations are prevalent**: Both intrinsic (manipulating input information) and extrinsic (adding new information) hallucinations occur frequently. Over 70% of single-sentence summaries generated by the studied systems contained hallucinations.
  "Intrinsic and extrinsic hallucinations happen frequently – in more than 70% of single-sentence summaries."
- **Most hallucinations are extrinsic and factually incorrect**: While extrinsic hallucinations could potentially be valid abstractions using background knowledge, over 90% of them were erroneous.
  "The majority of hallucinations are extrinsic, which potentially could be valid abstractions that use background knowledge. However, our study found that over 90% of extrinsic hallucinations were erroneous."
- **Pretrained models perform better**: Models initialized with pretrained parameters, such as BERTS2S, significantly outperformed other models in terms of faithfulness and factuality. This suggests that pretraining helps models integrate background knowledge and be less prone to generating incorrect information.
  "Third, models initialized with pretrained parameters perform best both on automatic metrics and human judgments of faithfulness/factuality. Furthermore, they have the highest percentage of extrinsic hallucinations that are factual."
- **Traditional metrics fail to capture faithfulness**: ROUGE and BERTScore, while useful for measuring informativeness, correlate poorly with human judgments of faithfulness and factuality.
  "Fourth, ROUGE (Lin and Hovy, 2003) and BERTScore (Zhang et al., 2020) correlates less with faithfulness/factuality than metrics derived from automatic semantic inference systems, specifically the degree to which a summary is entailed by the source document."
- **Semantic inference shows promise**: Textual entailment, which measures the degree to which a summary is entailed by the source document, shows stronger correlation with faithfulness and factuality. This highlights the potential for using semantic inference in automatic evaluation and even during model training and decoding.
  "This presents an opportunity for improved automatic evaluation measures as well as model training and decoding objectives."

Study Design:

- The study focused on the **extreme summarization task (XSUM)**, where models generate a single-sentence summary for a BBC news article.

- **Human evaluation** was central to the study. Annotators identified and categorized hallucinations, assessed their factuality and evaluated summaries for repetition and incoherence.
- **Five summarization systems** were compared: RNN-based Pointer-Generator (PTGEN), CNN-based Topic-aware Convolutional Seq2Seq (TCONVS2S), Transformer-based GPT-TUNED, TRANS2S, and BERTS2S, as well as the human-written reference summaries.

Implications:

- **Faithfulness and factuality are critical**: The study underscores the importance of addressing hallucinations in abstractive summarization.
- **Pretraining is crucial**: Pretraining models on large text corpora is essential for improving faithfulness and factuality.
- **New evaluation metrics are needed**: Traditional metrics like ROUGE are insufficient. Semantic inference-based metrics show promise for better capturing these crucial aspects of summarization quality.

Do the findings generalize to other datasets and types of summaries?

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks
https://paperswithcode.com/paper/retrieval-augmented-generation-for-knowledge#code
Source: Lewis, P. et al. (NeurIPS 2020).

Motivation: Large pre-trained language models store factual knowledge in their parameters, and achieve SOTA results when fine-tuned on downstream NLP tasks. However, their ability to access and precisely manipulate knowledge is still limited, and hence on knowledge-intensive tasks, their performance lags behind task-specific architectures. Additionally, providing provenance for their decisions and updating their world knowledge remain open research problems. Pre-trained models with a differentiable access mechanism to explicit non-parametric memory can overcome this issue, but have so far been only investigated for extractive downstream tasks. We explore a **general-purpose** fine-tuning recipe for retrieval-augmented generation (RAG) -- models which combine pre-trained parametric and non-parametric memory for language generation.

We compare two RAG formulations, one which conditions on the same retrieved passages across the whole generated sequence, the other can use different passages per token. We fine-tune and evaluate our models on a wide range of knowledge-intensive NLP tasks and set the state-of-the-art on three open domain QA tasks, outperforming parametric seq2seq models and task-specific retrieve-and-extract architectures. For language generation tasks, we find that RAG models generate more specific, diverse and factual language than a state-of-the-art parametric-only seq2seq baseline.

Hybrid language models: A novel approach to building language models that combine pre-trained **parametric memory** (e.g., BART seq2seq model) and **non-parametric memory** (e.g., a searchable index of Wikipedia).

- **RAG architecture:** RAG utilizes a **retriever** (DPR) to find relevant documents from the non-parametric memory based on the input. The retrieved documents are then fed to a **generator** (BART) along with the input to produce the output.
    - **RAG-Sequence:** Uses the same retrieved document for the entire output sequence.
    - **RAG-Token:** Allows different retrieved documents to contribute to each token in the output sequence.
    - **End-to-end training:** Both the retriever and generator are jointly trained by minimizing the negative marginal log-likelihood of target sequences.

From the Paper:

- "We explore a general-purpose fine-tuning recipe for retrieval-augmented generation (RAG) — models which combine pre-trained parametric and non-parametric memory for language generation."

- "RAG models achieve state-of-the-art results on open Natural Questions, WebQuestions and CuratedTrec and strongly outperform recent approaches that use specialised pre-training objectives on TriviaQA."
- "For language generation tasks, we find that RAG models generate more specific, diverse and factual language than a state-of-the-art parametric-only seq2seq baseline."
- "This example shows how parametric and non-parametric memories work together—the non-parametric component helps to guide the generation, drawing out specific knowledge stored in the parametric memory."

State-of-the-art results: RAG achieves new state-of-the-art results on open-domain question answering tasks like Natural Questions, WebQuestions, and CuratedTrec.

- **Knowledge-intensive tasks:** RAG excels at tasks requiring extensive knowledge, surpassing purely parametric models and specialized retrieval-based systems.
- **Improved generation:** RAG outperforms a strong BART baseline on abstractive QA (MS-MARCO) and Jeopardy question generation, producing more factual, specific, and diverse responses.
- **Competitive fact verification:** RAG achieves results close to state-of-the-art models on FEVER without requiring retrieval supervision.
- **Hot-swapping the index:** Demonstrated the ability to update the model's knowledge base by simply replacing the non-parametric memory, showing adaptability to changing information.

Benefits of Retrieval Augmentation: RAG leverages retrieval to access and manipulate knowledge more effectively, enabling:

- **Factual accuracy:** Reduced "hallucinations" compared to parametric-only models.
- **Specificity and Diversity:** More detailed and varied responses.
- **Knowledge update:** Easily update world knowledge by swapping the non-parametric memory index.
- **Interpretability:** Provides insights into model decisions through retrieved documents.

Discussion Points:

- Potential biases within the non-parametric memory source (e.g., Wikipedia) could influence RAG's output.
- Future work includes exploring joint pre-training of the retriever and generator, as well as investigating the interplay between parametric and non-parametric memory.