HEAD-QA v2: Expanding a Healthcare Benchmark for Reasoning

Alexis Correa-Guillén, Carlos Gómez-Rodríguez, David Vilares
Universidade da Coruña, CITIC
Departamento de Ciencias de la Computación y Tecnologías de la Información
Campus de Elviña s/n 15071, A Coruña, Spain
{alexis.cguillen@udc.es, carlos.gomez, david.vilares}@udc.es

Abstract

We introduce HEAD-QA v2, an expanded and updated version of a Spanish/English healthcare multiple-choice reasoning dataset originally released by Vilares and Gómez-Rodríguez (2019). The update responds to the growing need for high-quality datasets that capture the linguistic and conceptual complexity of healthcare reasoning. We extend the dataset to over 12,000 questions from ten years of Spanish professional exams, benchmark several open-source LLMs using prompting, RAG, and probability-based answer selection, and provide additional multilingual versions to support future work. Results indicate that performance is mainly driven by model scale and intrinsic reasoning ability, with complex inference strategies obtaining limited gains. Together, these results establish HEAD-QA v2 as a reliable resource for advancing research on biomedical reasoning and model improvement.

Keywords: Multi-choice question answering, LLMs, Healthcare

1. Introduction

HEAD-QA (v1) (Vilares and Gómez-Rodríguez, 2019) is a Spanish/English multiple-choice healthcare dataset designed to evaluate model reasoning abilities. It comprises 6,765 questions from official exams issued between 2013 and 2017. It was conceived as a step toward more demanding benchmarks, following the rise of early reading comprehension datasets such as SQuAD (Rajpurkar et al., 2016), SNLI (Bowman et al., 2015), and the AI2 Reasoning Challenge (Clark et al., 2018), among others, as well as the neural architectures developed for them (Kumar et al., 2016; Chen et al., 2017). Notably, experimental results revealed that these architectures lacked the capacity to reason effectively about diagnostic knowledge and failed to capture definitions and domain-specific concepts essential for accurate inference, often performing worse than simple information retrieval baselines.

More specifically, HEAD-QA consists of multiple-choice questions modeled after Spain's competitive specialization exams (Ministerio de Sanidad de España, 2023), which are used to evaluate and rank graduates in fields such as medicine (MIR), nursing (EIR), biology (BIR), chemistry (QIR), psychology (PIR), and pharmacy (FIR). These highly demanding exams require months or even years of preparation, as their results determine both the specialization and the training location where candidates complete the final 3–5 years of residency before becoming fully qualified professionals. The dataset has since gained notable adoption, having been used to evaluate influential architectures and models such as RMKV (Peng et al., 2023), Falcon (Penedo et al., 2023) and OLMo (Groeneveld et al., 2024), to investigate data reliability in both open-source and proprietary systems (Elazar et al., 2023), and to develop and assess specialized solutions in the medical domain (Zhang et al., 2023; Wang et al., 2024). It has also served as a precursor to similar medical QA datasets in other languages, including Chinese (Li et al., 2021) and French (Labrak et al., 2022), extending its influence on healthcare question answering research.

In the current context, the landscape of question answering and reasoning has changed profoundly with the rise of large language models (LLMs) (OpenAI, 2023; Jiang et al., 2024a; Dubey et al., 2024; Liu et al., 2024a; Gemma, 2025; Yang et al., 2025). These models have advanced substantially in reasoning, knowledge integration, and domain adaptation through instruction tuning and retrieval-augmented generation (RAG). This shift has redefined what constitutes a challenging benchmark—spanning domains such as coding (Zheng et al., 2025), Ph.D.-level knowledge (Phan et al., 2025), machine translation (Andrews et al., 2025) and multimodal reasoning (Padlewski et al., 2024)—and has led to an explosion of datasets (Rogers et al., 2023; Liu et al., 2024b).

Contribution We present HEAD-QA v2, an expanded and updated version designed to better reflect the era of large-scale reasoning models. The new release addresses the limited size and temporal coverage of its predecessor by incorporating 12,751 multiple-choice questions from Spanish professional medical qualification exams—more than doubling the dataset and extending its time span. We expect this expansion to enable future research on model generalization, knowledge reten-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5203–5214
11–16 May 2026.

tion, and temporal effects to a greater extent than its predecessor. We further establish new baselines through a systematic evaluation of open-source LLMs, exploring multiple inference strategies, including prompting, retrieval-augmented generation, and a probability-based approach. Together, we expect that these contributions offer a practical benchmark for studying how LLMs adapt to domain evolution, balance accuracy with efficiency, and perform complex reasoning in specialized contexts. The dataset is available at https://huggingface.co/datasets/alesi12/head_qa_v2.

2. Dataset Construction

This section outlines the construction of HEAD-QA v2, which, like its predecessor, is based on official, publicly available exams from the Ministerio de Sanidad de España. Each exam includes: (i) a two-column PDF containing the text, (ii) a CSV file listing the correct answers, and (iii) when applicable, a folder with referenced images indexed numerically (e.g., 1, 2, 3, 4, . . .), enabling text–image alignment.[1]

2.1. Preprocessing

The preprocessing pipeline involves converting, cleaning, and standardizing the exam data.

1. PDF to text conversion. Exams were converted from PDF to plain text using `pdftotext`, preserving the two-column layout.
2. Image mapping. Images were automatically linked, as related questions begin with "Question linked to image no. X," where X is the image identifier.
3. Question filtering. Questions without an official answer from the Spanish Ministry of Health were removed, as they correspond to disputed or withdrawn items.
4. Manual corrections. Minor edits to fix errors and standardize content. Chemical formulas were converted to SMILES notation (see Figure 1) using Mathpix , facilitating processing by text-based ML models (Schwaller et al., 2017; Chithrananda et al., 2020). The few affected questions were processed manually.

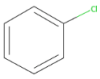

Figure 1: Example of chemical formula converted to SMILES notation for text processing.

[1]Questions containing images are relatively rare, and visual processing is therefore excluded from this work.

5. Storage. Files are stored in Parquet format (The Apache Software Foundation, 2024) for efficient compression and fast download.

2.2. Format

Each question includes eight fields (Figure 2). A unique identifier requires both `name` (exam name) and `qid` (question ID).

- `qid` (int): Question number within the exam.
- `qtext` (str): Question text.
- `ra` (int): Correct answer identifier.
- `answers` (list): Answer options, each with:
  - `aid` (int): Option ID.
  - `atext` (str): Option text.
- `image` (Image): Associated image in `PIL` format (Clark and Contributors, 2023), or `null` if none.
- `year` (int): Exam year.
- `category` (str): Discipline (e.g., Medicine, Nursing).
- `name` (str): Exam identifier combining year, discipline, and version (e.g., Cuaderno_2013_0_B).

```
{'qid': 1,
 'qtext': 'Excitatory postsynaptic
     potentials:',
 'ra': 3,
 'answers': [{'aid': 1, 'atext': 'Are all-
     or-none responses.'},
    {'aid': 2, 'atext': 'Are
        hyperpolarizing.'},
    {'aid': 3, 'atext': 'Can be summed.'},
    {'aid': 4, 'atext': 'Propagate over
        long distances.'},
    {'aid': 5, 'atext': 'Exhibit a
        refractory period.'}],
 'image': None,
 'year': 2013,
 'category': 'biology',
 'name': 'Cuaderno_2013_1_B'}
```

Figure 2: A HEAD-QA v2 question in JSON format.

2.3. Dataset statistics

The dataset contains a total of 12,751 questions distributed across six disciplines and ten years (see Table 1). Among them, 334 questions include images. Of these, 36 correspond to the four most recent nursing exams (2019–2022), while the rest belong to the medical exams—with over 30 image-based questions per test until 2018, and around 25 per test in subsequent years.

	'13	'14	'15	'16	'17	'18	'19	'20	'21	'22	Total
BIR	227	225	226	228	226	221	177	177	203	209	2 119
QIR	228	228	228	231	227	229	179	179	205	206	2 140
MIR	227	228	231	232	231	230	181	183	207	206	2 156
EIR	181	203	230	223	232	228	181	180	206	205	2 069
FIR	229	228	225	228	229	228	180	182	207	210	2 146
PIR	226	227	226	230	225	228	180	173	202	204	2 121
Total	1 318	1 339	1 366	1 372	1 370	1 364	1 078	1 074	1 230	1 240	12 751

Table 1: Number of questions per discipline/year.

Each question has one and only one correct answer. In the 2013 and 2014 exams, questions include five possible answers (2,657 items, representing 21% of the total), while the remaining exams feature four options per question. The correct answer is approximately uniformly distributed across the available options, although it is slightly less likely to appear in the first and last positions. This is not specific to this dataset but rather a well-documented bias in test design, as examiners tend to avoid placing the correct answer at the extremes (Attali and Bar-Hillel, 2003). This minor imbalance is not directly relevant to the purposes of this work, as no model is trained or conditioned on answer positions. Yet, recent studies have showed that LLMs exhibit positional biases in multiple-choice tasks, slightly favoring middle options (Pezeshkpour and Hruschka, 2023; Zheng et al., 2024).

In terms of question length (Figure 3), it remains stable over time, with the trend observed in HEAD-QA v1 persisting in recent years. Differences are more evident across disciplines (Figure 4): questions in biology, chemistry, pharmacology, and psychology tend to be shorter, while those in medicine and nursing are generally longer and detailed, often involving diagnostic reasoning that requires precise, context-rich information.

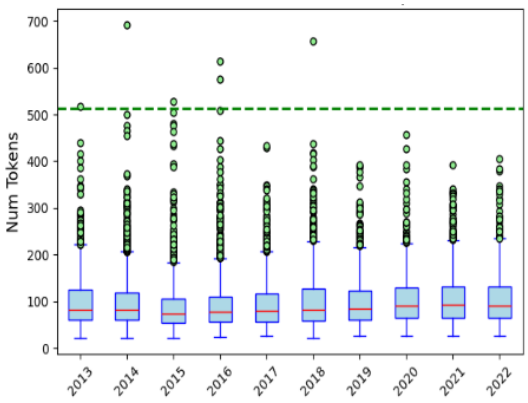

Figure 3: Question length distribution by year.

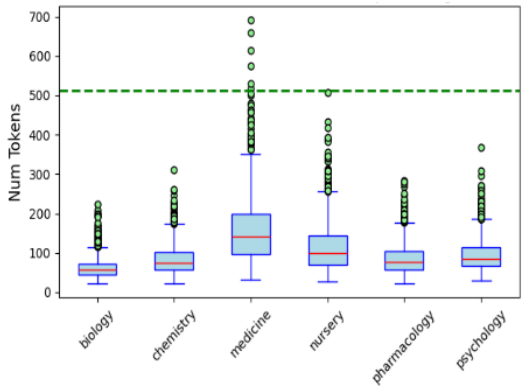

Figure 4: Question length distribution by discipline.

2.4. Machine Translation and Variants

To assess the impact of language variation, we consider the original Spanish dataset and its machine-translated English version, based on the approach of Vilares and Gómez-Rodríguez, who addressed the same objective in HEAD-QA v1 using Google's seq2seq model. For v2, we follow the recent trend of using LLMs for translation, leveraging their strong contextual reasoning and ability to process longer inputs while maintaining high translation quality across domains (Vilar et al., 2023; Zhu et al., 2024). In particular, we adopt LLaMA-3.1-8B and its instruction-tuned variant.

Translation prompt. We explored three prompting configurations: (i) zero-shot, providing a minimal translation instruction; (ii) one-shot, adding a single manually translated example mirroring the HEAD-QA format; and (iii) an instruction-tuned setup where the `system` message defines the model as an "expert translator," sets the Spanish→English direction, and enforces two rules: (a) preserve the multiple-choice format, and (b) output only the translation. The `user` message is the Spanish question and options verbatim.

Format integrity. To maintain structural parity with the source, we apply light post-processing: early stopping when the model emits the last option or begins a new prompt keyword (e.g., `SPANISH`); removal of trailing, non-requested text; normalization of option identifiers (replacing variants like "A)", "a.", etc., with `1.`, `2.`, ...); and validation checks to ensure that the output is a proper translation rather than an attempted answer or commentary (i.e., non-empty question and options, and consistent number of options). We check output validity through automatic checks for (i) empty text in the question or options, (ii) mismatched number of options, and (iii) incorrect numbering (e.g., repeated or misordered identifiers). For English, the instruction-tuned configuration produced the fewest errors (28), followed by the zero-shot (44) and one-shot (186) setups. This pattern shows that these models adhere to structured translation prompts, obtaining stable, well-aligned outputs.

Selection of final translations. Each question has three translated versions per target language, corresponding to different prompting configurations. The final dataset is compiled by selecting the most reliable translation according to the following rules: (i) if only one version is valid, it is kept; (ii) if several are valid, the one from the configuration with fewer errors is chosen; and (iii) if all are valid, the two most similar are compared, selecting the one from the lower-error setup. Questions without a valid translation are discarded. A manual evaluation of a random sample of questions was conducted to verify the reliability and consistency of this selection procedure.

Other language variants. Using the same translation and selection pipeline, we additionally generated Italian, Galician, and Russian versions of the dataset. Automatic format checks confirmed good structural consistency across these languages. While model evaluation was not conducted on these versions—since, unlike for English, no human validation could be performed in the final selection step due to resource constraints—they will be released alongside the main dataset to serve as a foundation for future research on cross-lingual and multilingual evaluation within the HEAD-QA framework.

Qualitative evaluation of translations. Still, to automatically assess the quality of the full translated datasets (English, Italian, Russian, and Galician) and enable comparison with future versions, we apply a back-translation (BT) approach. Each target-language version is translated back into Spanish and compared with the original text, obtaining round-trip translation (RTT) scores as a reference-free quality proxy (Zhuo et al., 2023). We compute both surface-level (BLEU) and semantic (BERTScore) similarity metrics. Results show that Galician and Italian achieve the highest BLEU (0.66 and 0.57) and BERTScore-F1 (0.80 and 0.77), followed by English (0.41 / 0.69) and Russian (0.33 / 0.65). These values are strong overall and consistent with linguistic distance, languages closer to Spanish yield higher lexical and semantic similarity, confirming that the translation pipeline maintains robust and semantically reliable outputs across all languages.

3. Baselines and Inference Strategies

Next, we present the models and inference strategies adopted, following standard practices.

3.1. Models

We evaluate four open-access, instruction-tuned LLMs:

Llama 3.1 (8B, 70B) (Dubey et al., 2024) Decoder-only models with 8B and 70B parameters, trained on multilingual data and officially supporting several languages beyond English, including Spanish. Both are optimized for long-context processing and use grouped-query attention (GQA) (Ainslie et al., 2023) to improve inference efficiency over standard multi-head attention (Vaswani et al., 2017).

Mistral v0.3 (7B) (Jiang et al., 2023) A 7B decoder-only model that combines grouped-query and sliding-window attention for efficient processing of sequences.

Mixtral v0.1 (8×7B) (Jiang et al., 2024b) Architecturally similar to Mistral 7B, Mixtral introduces a Mixture-of-Experts (MoE) design, activating two of eight experts per token to enhance efficiency by limiting active computation at each step.

The two model families, Llama 3.1 and Mistral, were selected for their broad adoption and good performance across diverse NLP tasks. Choosing one smaller and one larger model from each family enables a controlled examination of scaling effects in HEAD-QA v2, clarifying how model capacity influences biomedical reasoning and multiple-choice performance. While an exhaustive comparison across all available LLMs is beyond this study's scope, these models span both dense and mixture-of-experts architectures, offering a representative and methodologically sound basis for analysis. Since the primary objective of this work is the dataset itself, model evaluation serves mainly to characterize its difficulty and illustrate how different architectures respond to its challenges.[2]

3.2. Answer Selection Strategies

Each model answers multiple-choice questions using a consistent input–output scheme.

Model Input. By default, each question is formatted as a single text sequence that includes the question stem and its possible answers, each preceded by a numerical index, as illustrated in Figure 5.

Model Output. For all inference strategies, the model is queried to produce a short JSON structure indicating the index of the selected answer.

[2]All experiments were conducted under consistent hardware conditions using NVIDIA A100 GPUs (40 GB) with 16-bit precision. Smaller models (*Mistral-7B* and *Llama-3.1-8B*) were run on single-GPU nodes, whereas larger ones (*Llama-3.1-70B* and *Mixtral-8x7B*) required four GPUs, distributing the computational load evenly across devices.

```
Excitatory postsynaptic potentials:
1. Are all-or-none.
2. Are hyperpolarizing.
3. Can be summed.
4. Propagate over long distances.
5. Have a refractory period.
```

Figure 5: HEAD-QA v2 question encoded as a single input sequence.

For example, if the chosen option is the third, the expected output is `{answer: 3}`. Enforcing a fixed output format simplifies both extraction and post-processing of predictions, regardless of minor variations in spacing, casing, or punctuation.

3.2.1. Prompting Strategies

Zero-shot prompting Figure 6 shows the prompt used in the zero-shot setting. It defines the expected output format and provides minimal conditioning, instructing the model to act as an expert in scientific and healthcare domains.

```
<|begin_of_text|><|start_header_id|>system
    <|end_header_id|>

<|eot_id|><|start_header_id|>user<|
    end_header_id|>

You are an expert in specialized scientific
     and health disciplines. Respond to the
     following multiple-choice question:
Provide the answer in the following JSON
    format: {Answer: [number]}
For example, if the answer is 1, write: {
    Answer: 1}<|eot_id|><|start_header_id|>
    user<|end_header_id|>

<PLACEHOLDER FOR THE QUESTION AND OPTIONS
    ><|eot_id|><|start_header_id|>assistant
    <|end_header_id|>
```

Figure 6: Zero-shot prompt. The example, for Llama-3.1, shows the use of headers and special tokens that delimit user–assistant interactions and metadata as specified by the model architecture.

In-context learning LLMs often perform better when given examples within the prompt, as these help condition their responses. In this work, Figure 7 shows the few-shot prompt for Spanish questions, which includes three fixed examples from diverse disciplines. These examples, adapted from the United States Medical Licensing Examination (USMLE) questions, were selected to match the nature of HEAD-QA.[3] While a detailed analysis is beyond the scope of this study, prior work has shown that the choice and quality of in-context examples can strongly influence performance (Bonisoli et al., 2025). This phenomenon has also been interpreted as a form of implicit learning during inference (Dherin et al., 2025), suggesting that models may adapt dynamically—an ability that would be particularly relevant for sensitive (and very personalized) domains such as healthcare.

[3]Parallel Spanish and English versions were created to ensure linguistic and domain consistency.

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>

You are an expert in specialized scientific and health
      disciplines. Respond to the following multiple-choice
      question by indicating only the number of the correct
      option. No explanations are needed.<|eot_id|><|
      start_header_id|>user<|end_header_id|>

Which neurotransmitter is primarily involved in mood
      regulation?
1. Dopamine
2. Serotonin
3. GABA
4. Acetylcholine<|eot_id|><|start_header_id|>assistant<|
      end_header_id|>

{Answer: 2}<|eot_id|><|start_header_id|>user<|end_header_id
      |>

Which of the following is an example of a neutralization
      reaction in chemistry?
1. CH4 + 2O2 -> CO2 + 2H2O
2. Na + Cl2 -> 2NaCl
3. 2H2 + O2 -> 2H2O
4. HCl + NaOH -> NaCl + H2O<|eot_id|><|start_header_id|>
      assistant<|end_header_id|>

{Answer: 4}<|eot_id|><|start_header_id|>user<|end_header_id
      |>

...

<PLACEHOLDER FOR THE QUESTION AND OPTIONS><|eot_id|><|
      start_header_id|>assistant<|end_header_id|>
```

Figure 7: Example of a few-shot prompt with samples. Case shown for the Llama-3.1-8B model.

Chain-of-Thought prompting In this setting, the model is instructed to produce brief reasoning steps before providing an answer. As shown in Figure 8, the prompt asks the model to evaluate each option before selecting the most plausible one. This design encourages reasoning while keeping generations concise and inference efficient.

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>

You are an expert in scientific and health disciplines.
      Carefully analyze the following multiple-choice
      question and provide the correct answer. There is one
      and only one correct answer. Think through each option
      briefly before responding in the JSON format: {Answer:
      [number]}.<|eot_id|><|start_header_id|>user<|
      end_header_id|>

...
```

Figure 8: Example of a CoT prompt with brief reasoning before the final answer, using the Llama-3.1-8B model.

3.2.2. Retrieval-Augmented Generation

Following an approach shown to improve biomedical question answering (Xiong et al., 2024), in this work we also aim to mitigate potential hallucinations by retrieving relevant passages from an external, reliable corpus and appending them to the model's prompt to better guide its responses.

Our RAG implementation consists of three components: (i) an LLM, (ii) a biomedical corpus, and (iii) a retrieval system. For (i), we use the models introduced in §3.1. For (ii), we use the corpus proposed by Jin et al. (2021), which contains 18 medical textbooks commonly used for USMLE preparation.[4] For (iii), we use MedCPT (Jin et al., 2023), a dual-encoder model based on BERT (Devlin et al., 2019). It includes two specialized encoders—ncbi/MedCPT-Article-Encoder and ncbi/MedCPT-Query-Encoder—that map corpus fragments and queries (here, HEAD-QA v2 questions) into 768-dimensional vectors.[5,6]

Since the corpus is in English, retrieval was performed using English-translated versions of the questions, and the retrieved passages were reused for both the English and Spanish versions of the benchmark. Each question was paired with the two most relevant fragments, balancing contextual coverage with efficiency during LLM inference (together with the zero-shot prompt).

Assessing corpus alignment To evaluate the suitability of this corpus for our benchmark, Figures 9, 10, and 11 show a two-dimensional UMAP (McInnes et al., 2020) projection of all 126k corpus fragments and the 12k HEAD-QA v2 questions. Distinct clusters correspond to individual textbooks, with minimal overlap. Importantly, most HEAD-QA v2 questions project into high-density corpus regions, indicating strong topical alignment. For example, psychology questions cluster around Psychiatry_DSM-5 and Neurology_Adams, while biology and pharmacology items align with related sources. These observations suggest that the corpus and retrieval setup may supply relevant contextual evidence, motivating their inclusion as a baseline for our benchmark.

3.2.3. Selection via log-probabilities

Unlike the previous methods, which require autoregressive text generation, this approach directly compares the probabilities that a language model assigns to each candidate answer sequence.

Formally, let $C = (c_1, c_2, \ldots, c_n)$ represent the token sequence of a question and $A_i = (a_1, a_2, \ldots, a_m)$ the sequence corresponding to the i-th answer option. For each token a_j, the model computes a conditional probability $q_j = P(X_{n+j} \mid X_1 = c_1, \ldots, X_n = c_n, X_{n+1} = a_1, \ldots, X_{n+j-1} = a_{j-1})$. The overall likelihood of an answer sequence is then defined as the geometric mean of its token probabilities, $P(A_i) = \left(\prod_{j=1}^{m} q_j(a_j)\right)^{1/m}$. The model selects as correct the option that maximizes this probability, i.e., $= \arg\max_i P(A_i)$. Because multiplying many small probabilities can lead to numerical instability, all computations are performed in 32-bit precision. In addition, probabilities are evaluated in log-space to improve stability and efficiency, using the equivalent formulation $\log P(A_i) = \frac{1}{m} \sum_{j=1}^{m} \log q_j(a_j)$.

4. Experimental setup

Performance is evaluated using three metrics:(1) accuracy, the proportion of correct answers; (2) the normalized exam score, based on the official Spanish medical exam scheme (three wrong answers cancel one correct) and normalized by total items; and (3) the unanswered ratio, the fraction of questions with no valid response.

4.1. 'Prompting Strategy' Evaluation

Table 2 reports performance metrics for all prompting configurations (zero-shot, few-shot, and CoT).

Prompt	Model	English (en)			Spanish (es)		
		Acc	Score	P_{na}	Acc	Score	P_{na}
Zero-shot	Mixtral-8x7B	70.59	66.97	2.03	66.01	60.43	4.94
	Mistral-7B	59.55	52.61	4.82	52.79	43.56	3.25
	Llama-3.1-8B	70.43	67.86	0.39	61.93	56.61	0.38
	Llama-3.1-70B	**83.15**	84.16	0.43	**83.27**	84.14	0.40
Few-shot	Mixtral-8x7B	69.78	66.23	4.64	66.85	62.05	3.83
	Mistral-7B	60.63	54.59	4.42	54.06	45.90	3.02
	Llama-3.1-8B	70.49	68.24	**0.36**	62.58	57.44	**0.24**
	Llama-3.1-70B	82.90	84.14	0.51	83.24	**84.41**	0.38
CoT	Mixtral-8x7B	67.08	62.53	6.71	64.27	58.66	7.22
	Mistral-7B	56.19	47.89	8.54	48.07	36.55	10.55
	Llama-3.1-8B	69.13	66.50	5.55	61.05	55.30	6.11
	Llama-3.1-70B	82.54	**84.20**	2.11	82.10	83.21	4.07

Table 2: Performance metrics (accuracy, exam score, and proportion of unanswered questions) for all prompting configurations (zero-shot, few-shot, and CoT) in English and Spanish. Best values per column are highlighted in bold.

Overall, performance is consistently higher in English than in Spanish across all configurations, except for Llama-3.1-70B, where results are equivalent. This confirms that models handle English—either natively or through translation—more effectively. The gap is particularly pronounced

[4]This dataset, publicly available on the Hugging Face Hub (https://huggingface.co/datasets/MedRAG/textbooks), consists of approximately 126,000 short text fragments, each under 1,000 characters.

[5]Semantic similarity is computed via dot product. These models were trained on 255 million PubMed query–article pairs, making them highly effective for biomedical retrieval.

[6]For similarity search, we use FAISS (Facebook AI Similarity Search) (Douze et al., 2024), leveraging its native integration with the Hugging Face `datasets` library for low-memory data handling. We employ a flat index type, which performs exhaustive comparison across all vectors with 32-bit precision, ensuring maximal retrieval accuracy.

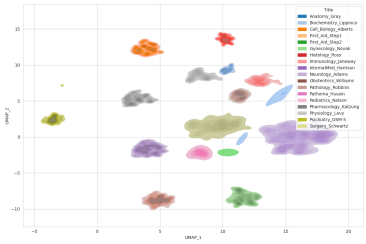

Figure 9: Kernel density estimation of corpus fragments by textbook source.

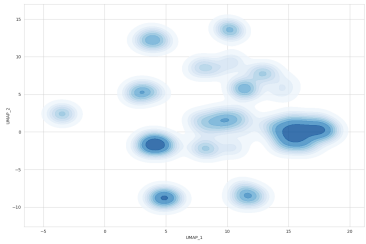

Figure 10: Global kernel density estimation of the corpus (without separating by textbook).

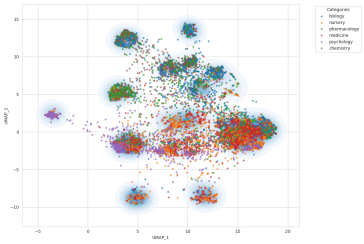

Figure 11: Scatter plot of HEAD-QA v2 questions by discipline overlaid on the corpus density map.

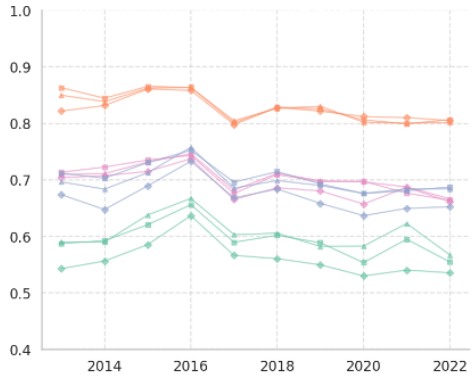

Figure 12: Performance evolution over time for each model on the English subset under the prompting setup. Colors indicate model families: Mistral-7B (green), Mixtral-8x7B (blue), Llama-3.1-70B (orange), and Llama-3.1-8B (pink). Markers denote prompting strategies: squares for zero-shot, triangles for few-shot, and diamonds for CoT.

in smaller models, suggesting that limited capacity (at least to represent specialized healthcare knowledge) amplifies cross-lingual variability. In contrast, larger models show stronger generalization, narrowing the difference between languages. Model scale has a clear impact: accuracy and exam scores increase steadily with model size, while the proportion of unanswered questions decreases.

Regarding prompting strategies, zero-shot and few-shot approaches achieve comparable results, suggesting that providing a single example offers limited additional benefit given the models' instruction tuning. Exploring the impact of example selection could be an interesting direction for future work. In contrast—perhaps unexpectedly—CoT prompting consistently reduces accuracy and increases non-response rates except for the Llama-3.1-70B, indicating that explicit reasoning steps may actually reduce performance in this healthcare domain.

Figure 12 shows that English performance remains stable across exam years, with larger models outperforming smaller ones. English results are slightly higher than Spanish (not shown for space), and simpler prompting strategies get the most reliable outcomes.

4.2. 'RAG Strategy' Evaluation

Prompt	Model	English (en)			Spanish (es)		
		Acc	Score	P_{na}	Acc	Score	P_{na}
RAG	Mixtral-8x7B	69.80	66.25	4.67	66.90	62.07	3.91
	Mistral-7B	56.63	49.11	2.14	49.61	39.70	2.30
	Llama-3.1-8B	66.45	62.83	0.69	59.13	52.86	0.57
	Llama-3.1-70B	**82.45**	**83.13**	**0.32**	**82.52**	**83.03**	**0.22**

Table 3: Performance metrics (accuracy, exam score, and proportion of unanswered questions) for all RAG-based configurations in English and Spanish.

Table 3 presents the performance metrics for the models that used RAG to condition their prompt.

Overall, results show that incorporating retrieved context through RAG does not lead to consistent improvements over standard prompting. Performance remains slightly higher in English than in Spanish. Larger models benefit the most from RAG, maintaining competitive accuracy and lower non-response rates, while smaller models tend to degrade when exposed to noisy or weakly relevant evidence.

Compared to the zero-shot baseline, RAG gets slightly lower scores in both languages. This suggests that the retrieved passages are not always effectively integrated into the generation process, that the model can often rely on its internal knowledge instead, or that the retrieved information is not sufficiently relevant. The weak correlation between retrieval relevance (see §3.2.2) and answer correctness ($r = 0.07$) further supports this interpretation: model performance appears to depend primarily on internal knowledge rather than external evidence. Yearly trends remain stable across models and languages, closely mirroring those observed in the prompting experiments, as shown in Figure 13.

4.3. 'Log-probability' Evaluation

Table 4 reports accuracy and normalized exam scores for this setup. By design, the unanswered

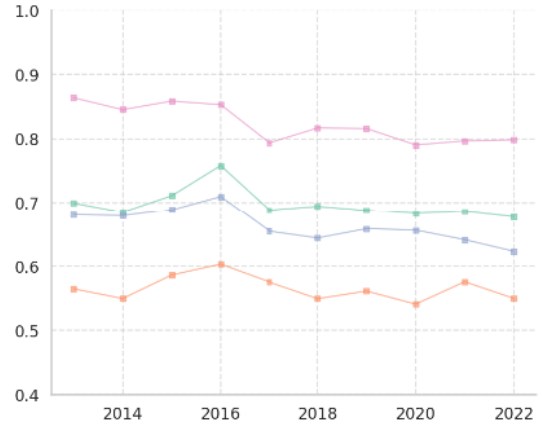

Figure 13: Performance evolution over time for each model in English under the RAG setup. Colors indicate model families: Mixtral-8x7B (green), Mistral-7B (orange), Llama-3.1-8B (blue), and Llama-3.1-70B (pink).

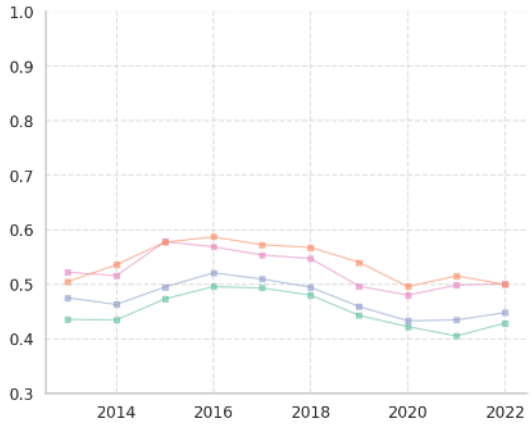

Figure 14: Performance evolution over time for each model in English under the probability-based selection setup. Colors indicate model families: Llama-3.1-8B (green), Llama-3.1-70B (orange), Mistral-7B (blue), and Mixtral-8x7B (pink).

ratio is 0%, as models are required to select one option per question. Despite this, scores drop notably compared to prompting-based approaches, indicating that direct likelihood evaluation is less effective for multiple-choice reasoning. Performance remains consistently higher in English than in Spanish, with the gap being more pronounced in smaller models. Larger models mitigate this difference, maintaining more stable accuracy across languages. The observed performance gap can also be attributed to the independent evaluation of each option: without jointly considering all alternatives, models lose the elimination-based reasoning that benefits prompting approaches.

As shown in Figure 14, yearly trends remain stable, with only minor fluctuations after 2018. Although this method minimizes resource usage—since no text generation is involved—the efficiency gain does not compensate for the accuracy loss, limiting its practical value for HEAD-QA v2.

Strategy	Model	English (en)		Spanish (es)	
		Acc	Score	Acc	Score
Prob-based	Mixtral-8x7B	52.84	44.06	47.87	37.60
	Mistral-7B	47.86	37.49	39.42	27.69
	Llama-3.1-8B	45.25	34.31	37.82	24.80
	Llama-3.1-70B	**54.15**	**46.04**	**51.42**	**42.12**

Table 4: Performance metrics (accuracy and exam score) for the probability-based selection strategy (Section 3.2.3) in English and Spanish.

5. Discussion

The results reveal consistent trends across model families, highlighting how architectural scale, language, and methodological design shape performance in HEAD-QA v2. Model size emerges as the most decisive factor: Llama-3.1-70B consistently achieves the highest accuracy and normalized exam scores, while smallest models performs lowest across all metrics. These results align with broader findings in LLM evaluation, where scaling enhances both factual recall and reasoning stability.

Language effects are present but moderate, with smaller models showing slightly reduced performance in Spanish. This may stem from differences in tokenization efficiency, knowlegde integration, and from weaker multilingual representations in smaller architectures, which maybe be less robust to lexical and syntactic variability across languages.

Methodologically, neither more elaborate prompting (few-shot or CoT) nor retrieval-augmented generation produces consistent improvements. In some cases, these strategies even reduce performance, suggesting that additional contextual input can introduce noise or divert the model from leveraging its internal knowledge. Considering their higher computational and developmental costs, such methods offer limited benefit in this setting.

Finally, the probability-based answer selection strategy performs notably worse than generation-based approaches. Since each option is scored independently, the model cannot perform the comparative reasoning and contextual alignment typical of human multiple-choice problem-solving, resulting in systematic accuracy drops.

6. Conclusion

This work introduced HEAD-QA v2, a new large-scale, multilingual benchmark designed to evaluate complex reasoning in the biomedical domain. Through extensive experiments across multiple modern large language models and inference strategies, we established empirical baselines and

analyzed the factors that most influence performance. Our findings indicate that, for highly specialized biomedical question answering, the intrinsic knowledge and reasoning capacity of the language model play a far greater role than the sophistication of the inference strategy. Techniques such as RAG and CoT prompting, while successful in other domains, did not obtain consistent gains in this setting and introduced additional computational and implementation overhead. Overall, improvements on HEAD-QA v2 seem more closely tied to scaling and refining the underlying models than to increasing inference complexity, though alternative strategies may still offer potential for future exploration.

Limitations

This study did not include evaluations with frontier proprietary LLMs such as GPT-4, Claude, or Gemini, primarily due to funding resources to access APIs. Consequently, the results reflect trends among open-access models up to 70B parameters.

Additionally, while the English translations were automatically generated and reviewed for terminological consistency, large-scale human validation was not feasible. Minor translation inconsistencies could therefore influence model performance, especially in domain-specific terminology.

Another limitation concerns the scope of the benchmark itself. HEAD-QA v2 focuses on multiple-choice biomedical questions, which represent only a subset of complex reasoning skills.

Ethical Considerations

HEAD-QA v2 is based on publicly available examination questions designed for healthcare education, containing no personal or patient data. Nevertheless, the dataset and experiments involve content related to medical knowledge, and outputs from large language models should not be interpreted as clinical advice.

All experiments were conducted with open-access models and publicly available data, ensuring reproducibility and compliance with data use terms. We acknowledge that automatic translation and model-generated text may propagate biases or inaccuracies, and encourage caution when using the dataset or models in real-world or educational healthcare contexts.

Acknowledgments

We acknowledge grants GAP (PID2022-139308OA-I00) funded by MICIU/AEI/10.13039/501100011033/ and ERDF, EU; LATCHING (PID2023-147129OB-C21) funded by MICIU/AEI/10.13039/501100011033 and ERDF, EU; and TSI-100925-2023-1 funded by Ministry for Digital Transformation and Civil Service and "NextGenerationEU" PRTR; as well as funding by Xunta de Galicia (ED431C 2024/02), and CITIC, as a center accredited for excellence within the Galician University System and a member of the CIGUS Network, receives subsidies from the Department of Education, Science, Universities, and Vocational Training of the Xunta de Galicia. Additionally, it is co-financed by the EU through the FEDER Galicia 2021-27 operational program (Ref. ED431G 2023/01). This research project was made possible through the access granted by the Galician Supercomputing Center (CESGA) to its supercomputing infrastructure. The supercomputer FinisTerrae III and its permanent data storage system have been funded by the NextGeneration EU 2021 Recovery, Transformation and Resilience Plan, ICT2021-006904, and also from the Pluriregional Operational Programme of Spain 2014-2020 of the European Regional Development Fund (ERDF), ICTS-2019-02-CESGA-3, and from the State Programme for the Promotion of Scientific and Technical Research of Excellence of the State Plan for Scientific and Technical Research and Innovation 2013-2016 State subprogramme for scientific and technical infrastructures and equipment of ERDF, CESG15-DE-3114. Finally, we are grateful to the funding of *Consellería de Educación, Ciencia, Universidades e Formación Profesional (Xunta de Galicia - Convenio para o desenvolvemento de accións estratéxicas de I+D+i 2025-2026).*

7. Bibliographical References

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. Gqa: Training generalized multi-query transformer models from multi-head checkpoints.

Pierre Andrews, Mikel Artetxe, Mariano Coria Meglioli, Marta R Costa-jussà, Joe Chuang, David Dale, Cynthia Gao, Jean Maillard, Alex Mourachko, Christophe Ropers, et al. 2025. Bouquet: dataset, benchmark and open initiative for universal quality evaluation in translation. *arXiv preprint arXiv:2502.04314*.

Yigal Attali and Maya Bar-Hillel. 2003. Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. *Journal of Educational Measurement*, 40(2):109–128.

Giovanni Bonisoli, David Vilares, Federica Rollo,

and Laura Po. 2025. Document-level event extraction from italian crime news using minimal data. *Knowledge-Based Systems*, 317:113386.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2020. Chemberta: Large-scale self-supervised pretraining for molecular property prediction.

Alex Clark and Contributors. 2023. Pillow (pil fork). `https://python-pillow.org/`. Version 10.0.0, Accessed: October 13, 2025.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Benoit Dherin, Michael Munn, Hanna Mazzawi, Michael Wunder, and Javier Gonzalvo. 2025. Learning without training: The implicit dynamics of in-context learning.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, et al. 2023. What's in my big data? *arXiv preprint arXiv:2310.20707*.

Gemma. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. 2024. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024a. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024b. Mixtral of experts.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021.

What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11).

Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. 2016. Ask me anything: Dynamic memory networks for natural language processing. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1378–1387, New York, New York, USA. PMLR.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Beatrice Daille, Pierre-Antoine Gourraud, Emmanuel Morin, and Mickael Rouvier. 2022. FrenchMedMCQA: A French multiple-choice question answering dataset for medical domain. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 41–46, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jing Li, Shangping Zhong, and Kaizhi Chen. 2021. MLEC-QA: A Chinese Multi-Choice Biomedical Question Answering Dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8862–8874, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Yang Liu, Jiahuan Cao, Chongyu Liu, Kai Ding, and Lianwen Jin. 2024b. Datasets for large language models: A comprehensive survey. *arXiv preprint arXiv:2402.18041*.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction.

Ministerio de Sanidad de España. 2023. Cuadernos de examen - formación sanitaria especializada. https://fse.mscbs.gob.es/fseweb/view/public/datosanteriores/cuadernosExamen/busquedaConvocatoria.xhtml. Accedido el 19 de octubre de 2023.

OpenAI. 2023. Chatgpt (3.5 version). https://chat.openai.com/. Large language model.

Piotr Padlewski, Max Bain, Matthew Henderson, Zhongkai Zhu, Nishant Relan, Hai Pham, Donovan Ong, Kaloyan Aleksiev, Aitor Ormazabal, Samuel Phua, et al. 2024. Vibe-eval: A hard evaluation suite for measuring progress of multimodal language models. *arXiv preprint arXiv:2405.02287*.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data only. In *Advances in Neural Information Processing Systems*, volume 36, pages 79155–79172. Curran Associates, Inc.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Leon Derczynski, Xingjian Du, Matteo Grella, Kranthi Gv, Xuzheng He, Haowen Hou, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartłomiej Koptyra, Hayden Lau, Jiaju Lin, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Guangyu Song, Xiangru Tang, Johan Wind, Stanisław Woźniak, Zhenyuan Zhang, Qinghua Zhou, Jian Zhu, and Rui-Jie Zhu. 2023. RWKV: Reinventing RNNs for the transformer era. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14048–14077, Singapore. Association for Computational Linguistics.

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. 2025. Humanity's last exam. *arXiv preprint arXiv:2501.14249*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *ACM Computing Surveys*, 55(10):1–45.

Philippe Schwaller, Theophile Gaudin, David Lanyi, Costas Bekas, and Teodoro Laino. 2017. "found in translation": Predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models.

The Apache Software Foundation. 2024. Apache parquet documentation. Accedido el 19 de octubre de 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

David Vilares and Carlos Gómez-Rodríguez. 2019. HEAD-QA: A healthcare dataset for complex reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 960–966, Florence, Italy. Association for Computational Linguistics.

Xidong Wang, Nuo Chen, Junyin Chen, Yidong Wang, Guorui Zhen, Chunxian Zhang, Xiangbo Wu, Yan Hu, Anningzhe Gao, Xiang Wan, et al. 2024. Apollo: A lightweight multilingual medical llm towards democratizing medical ai to 6b people. *arXiv preprint arXiv:2403.03640*.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. 2023. Alpacare: Instruction-tuned large language models for medical application. *arXiv preprint arXiv:2310.14558*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors.

Zihan Zheng, Zerui Cheng, Zeyu Shen, Shang Zhou, Kaiyuan Liu, Hansen He, Dongruixuan Li, Stanley Wei, Hangyi Hao, Jianzhu Yao, et al. 2025. Livecodebench pro: How do olympiad medalists judge llms in competitive programming? *arXiv preprint arXiv:2506.11928*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

Terry Yue Zhuo, Qiongkai Xu, Xuanli He, and Trevor Cohn. 2023. Rethinking round-trip translation for machine translation evaluation.