Table 2: **Evaluation on LitQA**. We compare PaperQA with other LLMs, the AutoGPT agent, and commercial products that use RAG. AutoGPT was run with GPT-4, where other implementation details are given in the Appendix B. Elicit.AI was run on default settings, Perplexity was run in academic mode, Perplexity Co-pilot was run on default settings (perplexity model, "all sources"), and "Assistant by Scite_" was run on default settings. Each question was run on a new context (thread) and all commercial products were evaluated on September 27, 2023. We report averages over a different number of runs for each.

Model	Samples	Response			Score	
		Correct	Incorrect	Unsure	Accuracy ($\frac{\text{Correct}}{\text{All}}$)	Precision ($\frac{\text{Correct}}{\text{Sure}}$)
Random	100	10.2	29.5	10.3	20.4%	25.7%
Human	5	33.4	4.6	12.0	66.8%	**87.9%**
Claude-2	3	20.3	26.3	3.3	40.6%	43.6%
GPT-4	3	16.7	16.3	17.0	33.4%	50.6%
AutoGPT	3	20.7	7.3	22.0	41.4%	73.9%
Elicit	1	12.0	16.0	22.0	24.0%	42.9%
Scite_	1	12.0	21.0	17.0	24.0%	36.4%
Perplexity	1	9.0	10.0	31.0	18.0%	47.4%
Perplexity (Co-pilot)	1	29.0	10.0	12.0	58.0%	74.4%
PaperQA	4	34.8	4.8	10.5	**69.5%**	**87.9%**

and Perplexity – in Table 2. All commercial tools are specifically tailored to answering questions by retrieving scientific literature. We give them the same prompt as to PaperQA. From Table 2 we see that PaperQA outperforms all competing models and products, and is on par with that of human experts with access to the internet. Furthermore, we see the lowest rate of incorrectly answered questions out of all tools, which rivals that of humans. This emphasizes that PaperQA is better calibrated to express uncertainty when it actually is uncertain. Surprisingly, GPT-4 and Claude-2 perform better than random although the questions are from papers after their training cut-off date, suggesting they have latent knowledge, leading to useful bias towards answers that are more plausible.

PaperQA averaged 4,500 tokens (prompt + completion) for the more expensive LLMs (`agent LLM`, `answer LLM`, `ask LLM`) and 24,000 tokens for the cheaper, high-throughput LLM (`summary LLM`). Based on commercial pricing as of September 2023, that gives a cost per question of $0.18 using the stated GPT-4 and GPT-3.5-turbo models. It took PaperQA on average about 2.4 hours to answer all questions, which is also on par with humans who were given 2.5 hours. A single instance of PaperQA would thus cost $3.75 per hour, which is a fraction of an average hourly wage of a desk researcher. We exclude other negligible operating costs, such as search engine APIs, or electricity.

How does PaperQA compare to expert humans? PaperQA shows similar results to those of the expert humans who answered the questions. To quantify this, we calculate the categorical correlation (Cramer's V) of the responses for each human-human and human-PaperQA pair. Average human-human V is 0.66 ± 0.03, whereas average human-PaperQA V is 0.67 ± 0.02 (mean ± stderr), indicating that PaperQA is, on average, as correlated with human respondents as the human respondents are with each other, implying no discernable difference in responses. To compare, the average V between humans and Perplexity was 0.630 ± 0.05.

Ablating PaperQA We report performance on LitQA when toggling different parts and LLMs of PaperQA in Table 3. Using GPT-4 as the `answer LLM` slightly outperforms Claude-2. When we look at the different components of PaperQA, we observe a major drop in performance when not including multiple-choice options as answers (*no MC options*) and using *Semantic Scholar* instead of Google Scholar. The former we explain with the fact that closed-form questions are easier than open-form ones, and the model can use keywords derived from the possible answers to search. The drop in performance of the linear settings, *Vanilla RAG* and *No search*, show the advantage of an agent-based model that can call tools multiple times until it is satisfied with the final answer. Surprisingly enough, not using the LLM's latent knowledge (*no ask LLM*) also hurts performance, despite the benchmark being based on information after the cutoff date – we suggest that the useful latent knowledge we find LLMs to possess in Table 5 helps the agent use the best pieces of evidence.