You are an expert superforecaster, familiar with the work of Tetlock and others. Make a prediction of the probability that the question will be resolved as true. You MUST give a probability estimate between 0 and 1 UNDER ALL CIRCUMSTANCES. If for some reason you can't answer, pick the base rate, but return a number between 0 and 1.

Question: {question}
Question Background: {background}
Resolution Criteria: {resolution_criteria}

Today's date: {date_begin}
Question close date: {date_end}

Output your answer (a number between 0 and 1) with an asterisk at the beginning and end of the decimal. Do not output anything else.
Answer: {{ Insert answer here }}

Figure 5: **The simple zero-shot prompt used for baseline evaluations**. No retrieval is performed. The prompt simply asks the model to make a prediction on a given question from the test set. We add the directive "You MUST ... UNDER ALL CIRCUMSTANCES" to push the model to answer the question, which in some cases it refuses to, potentially due to safety training. See Section 3.4 for results and Appendix B for more details.

Question: {question}
Question Background:{background}
Resolution Criteria:{resolution_criteria}

Today's date: {date_begin}
Question close date: {date_end}

Instructions:
1. Provide reasons why the answer might be no.
{{ Insert your thoughts }}

2. Provide reasons why the answer might be yes.
{{ Insert your thoughts }}

3. Aggregate your considerations.
{{ Insert your aggregated considerations }}

4. Output your answer (a number between 0 and 1) with an asterisk at the beginning and end of the decimal.
{{ Insert your answer }}

Figure 6: **The scratchpad prompt used for baseline evaluations.** No retrieval is performed. The prompt asks the model to make a prediction on a given question from the test set, after making considerations for yes and no. See Section 3.4 for results and Appendix B for more details.

B.2 Baseline Evaluation Results

We now give the full results of our baseline evaluation (Section 3.4) in Table 7.

Model	Zero-shot	Scratchpad
GPT-3.5-Turbo	0.237 (0.014)	0.257 (0.009)
GPT-3.5-Turbo-1106	0.274 (0.016)	0.261 (0.010)
GPT-4 (GPT-4-0613)	0.219 (0.013)	0.222 (0.009)
GPT-4-1106-Preview	**0.208** (0.013)	**0.209** (0.012)
Llama-2-7B	0.353 (0.020)	0.264 (0.011)
Llama-2-13B	0.226 (0.009)	0.268 (0.008)
Llama-2-70B	0.283 (0.014)	0.282 (0.011)
Mistral-7B-Instruct	0.237 (0.018)	0.243 (0.008)
Mistral-8x7B-Instruct	0.238 (0.018)	0.238 (0.010)
Mixtral-8x7B-DPO	0.260 (0.022)	0.248 (0.010)
Yi-34B-Chat	0.238 (0.012)	0.241 (0.009)
Claude-2	0.220 (0.013)	0.219 (0.014)
Claude-2.1	0.220 (0.013)	0.215 (0.014)
Gemini-Pro	0.243 (0.019)	0.230 (0.007)

Table 7: **Zero-shot and scratchpad Brier scores** on the test set: Brier scores under zero-shot or scratchpad prompts, with 2 standard error (SE) values. Lower is better. Random baseline: 0.250; human crowd: 0.149. All models fall significantly far from human aggregate.

B.3 Knowledge Evaluation by Category

We present an evaluation of model's knowledge about resolved questions on past events and notice variations in performance across categories. To investigate further, we analyzed each model's zero-shot Brier score on the test set by category. This analysis showed a correlation between models' knowledge on the training and validation sets and their Brier scores on the test set across categories. This suggests that domain-adaptive training could be used to improve model performance in categories where its existing knowledge is limited.

First, we assessed pre-trained language model knowledge across categories by evaluating their ability to answer resolved forecasting questions from the train and validation sets. See Table 8 for the results and Figure 7 for the knowledge prompt.

Model	Arts & Recreation	Economics & Business	Education & Research	Environment & Energy	Healthcare & Biology	Politics & Governance	Science & Tech	Security & Defense	Social Sciences	Sports	Other
GPT-3.5-Turbo	0.411	0.323	0.25	0.314	0.419	0.328	0.387	0.314	0.462	0.365	0.107
GPT-3.5-Turbo-1106	0.196	0.195	0.375	0.229	0.262	0.278	0.247	0.286	0.154	0.166	0.25
GPT-4 (GPT-4-0613)	0.083	0.114	0.125	0.22	0.157	0.349	0.196	0.279	0.077	0.04	0.071
GPT-4-1106-Preview	0.094	0.142	0.125	0.153	0.144	0.391	0.227	0.207	0.0	0.234	0.0
Llama-2-7B	0.042	0.069	0.156	0.203	0.284	0.046	0.067	0.033	0.0	0.05	0.071
Llama-2-13B	0.102	0.181	0.156	0.288	0.288	0.21	0.247	0.163	0.231	0.189	0.036
Llama-2-70B	0.143	0.175	0.344	0.322	0.266	0.243	0.384	0.115	0.077	0.075	0.107
Mistral-7B-Instruct	0.011	0.024	0.0	0.034	0.022	0.05	0.054	0.007	0.0	0.018	0.0
Mistral-8x7B-Instruct	0.019	0.045	0.094	0.051	0.066	0.071	0.049	0.04	0.0	0.027	0.0
Mixtral-8x7B-DPO	0.004	0.084	0.031	0.051	0.087	0.041	0.054	0.014	0.0	0.01	0.0
YI-34B-Chat	0.423	0.552	0.625	0.593	0.555	0.63	0.588	0.738	0.538	0.624	0.536
Claude-2	0.14	0.205	0.219	0.254	0.245	0.446	0.296	0.134	0.154	0.392	0.071
Claude-2.1	0.136	0.205	0.219	0.246	0.249	0.446	0.294	0.136	0.077	0.395	0.071
Gemini-Pro	0.155	0.425	0.188	0.415	0.314	0.425	0.356	0.545	0.077	0.35	0.25

Table 8: **Comparison of knowledge accuracy across categories and models** on the train and validation sets. We list the knowledge accuracy of all base models with respect to all categories in the train and validation set.

We noticed variations in knowledge accuracy across categories. To dig deeper, we analyze the zero-shot Brier score on the test set in Table 9 and assess if there is a correlation between knowledge accuracy on the training and validation sets and zero-shot Brier score on the test set in Table 10.

The potential for domain-adaptive training. We calculate the correlation between the models' knowledge accuracy and their Brier scores of the zero-shot evaluation. Notably, in the Politics & Governance, Arts & Recreation, and Education & Research categories, there exists a strong negative correlation. See the below

Question: {question}
The question was posed on {date_begin} and closed on {date_end}.
Instructions:

- Please output "1" if the answer is "Yes", "0" if the answer is "No" or "IDK" if you don't know the answer. Do not return anything else.
- Do not guess.

Answer: {{ Insert answer here }}

Figure 7: **The prompt used for evaluating model's knowledge about forecasting questions**. It asks the model to answer "Yes" or "No" given its pre-training knowledge and also allows for "IDK" ("I don't know"). See Section B.3 for the results.

Model	Arts & Recreation	Economics & Business	Education & Research	Environment & Energy	Healthcare & Biology	Politics & Governance	Science & Tech	Security & Defense	Sports
GPT-3.5-Turbo	0.292	0.281	0.270	0.245	0.388	0.244	0.178	0.235	0.205
GPT-3.5-Turbo-1106	0.309	0.294	0.336	0.239	0.336	0.343	0.225	0.250	0.214
GPT-4 (GPT-4-0613)	0.278	0.260	0.437	0.201	0.203	0.228	0.200	0.224	0.178
GPT-4-1106-Preview	0.240	0.244	0.394	0.222	0.122	0.218	0.178	0.207	0.177
Llama-2-7B	0.381	0.356	0.331	0.359	0.351	0.399	0.351	0.288	0.327
Llama-2-13B	0.260	0.247	0.263	0.218	0.230	0.245	0.197	0.222	0.199
Llama-2-70B	0.318	0.329	0.319	0.299	0.498	0.329	0.308	0.264	0.212
Mistral-7B-Instruct	0.291	0.265	0.295	0.228	0.238	0.271	0.184	0.236	0.191
Mistral-8x7B-Instruct	0.354	0.272	0.452	0.256	0.335	0.252	0.176	0.227	0.189
Mixtral-8x7B-DPO	0.367	0.315	0.543	0.213	0.217	0.287	0.184	0.265	0.194
YI-34B-Chat	0.263	0.240	0.332	0.196	0.208	0.265	0.196	0.236	0.212
Claude-2	0.293	0.239	0.326	0.199	0.226	0.214	0.175	0.244	0.194
Claude-2.1	0.293	0.242	0.316	0.199	0.226	0.213	0.183	0.244	0.194
Gemini-Pro	0.301	0.303	0.432	0.227	0.210	0.263	0.175	0.255	0.189

Table 9: **Comparison of zero-shot Brier scores across categories and models** on the test set. This table lists the Brier scores of all base models with respect to the specified categories.

Table 10 for the correlation table. This negative correlation is expected because a higher knowledge accuracy should intuitively correspond to a lower Brier score. As a direction for future research, we propose that domain-adaptive training could be employed to enhance forecasting performance in specific categories.

Category	Score
Arts & Recreation	**-0.417103**
Economics & Business	-0.228040
Education & Research	**-0.359102**
Environment & Energy	-0.135552
Healthcare & Biology	0.162110
Politics & Governance	**-0.487266**
Science & Tech	-0.091878
Security & Defense	-0.183253
Sports	-0.136017

Table 10: **Correlation between knowledge accuracy and zero-shot prompt Brier score by category**. Categories with an absolute correlation of 0.3 or greater, shown in bold, indicate a high correlation between accuracy on the training and validation set and forecasting performance on the test set. This highlights that in certain domains model's forecasting capabilities are correlated with its pre-training knowledge.