
Supplementary Materials

Anonymous Author(s)

Affiliation

Address

email

A Probe

Datasets. We run four probe experiments, each with its own dataset. For the layer-wise “dead / alive” probe, we compile 1,000 dead and 1,000 alive individuals. We split 80%/20% for training and testing. The second experiment evaluates whether the model knows death status for a specific year. We train separate probes for 60 reference years spanning 30 years before to 30 years after each subject’s death, labeling before-death years alive and after-death years dead. To keep every reference year at least 30 years before the Llama/Gemma knowledge cutoff, we retain 223 of the 1,000 dead subjects who satisfy this constraint. Each subject provides 30 question pairs, yielding 466 examples per reference year, which we split 80%/20% for training and testing. We include the hyperparameters used for these two type of probes in Table 1.

For temporal representation probes, we train on president-related data and more generalized data. We train temporal-representation probes on 277 U.S. president questions with an 80%/20% train–test split. In the ROLE-PLAY setting, we sample characters from our real-person dataset, partition them 80%/20%, and pair training (test) questions only with training (test) characters. For generalized temporal question probes, we apply the same procedure to the entertainment dataset proposed before [1], which contains 31,321 items (24,884 train, 6,437 test). We follow their original training protocol.

Additional probing results. In Section 4.1, we show that Llama lacks a reliable linear representation of death year. Gemma behaves similarly, as shown in Figure 1. For the “dead / alive” probe, Gemma’s accuracy gap between ROLE-PLAY and NON-ROLE-PLAY is smaller, but accuracy still drops in the ROLE-PLAY setting, indicating the representation is weaker. Year-specific death-status probes give the same result. Gemma, like Llama, fails to encode linearly separable representations of this information. To determine whether the model encodes a non-linear representation of the general “dead / alive” status or of year-specific death information, we train multilayer perceptron (MLP) probes on the same dataset used for the linear probes. As shown in Figure 2, the MLP probes do not outperform the linear ones, indicating that neither the binary life status nor the exact death year is captured nonlinearly in the hidden activations.

Section 4.2 reports temporal-representation deviations in the ROLE-PLAY setting, and Figure 3 adds Spearman correlations and RMSEs for both question sets in Llama and Gemma across all settings.

Table 2 reports Spearman correlations for general temporal questions binned into five-year chunks under the NON-ROLE-PLAY setting. Truncating to this smaller scale yields correlations below 0.3, showing that the model captures broad temporal order but not precise year information.

B Additional Results on Improved Specification

In Section 5, we enhance the after-death abstention and answer behaviors of Llama and Gemma, then replicate the same experiments on GPT and Claude. Figure 4 reveals the same pattern: abstention and answer behaviors improve, but accuracy drops. After-death abstention rises by 88.4% for Claude and 87.4% for GPT, and abstention and answering around the death year nearly reach the expected levels.

Table 1: Hyperparameters for probe training.

Hyperparameters	Dead/Alive	Death Year
learning rate	0.001	0.001
batch size	100	100
epochs	500	500

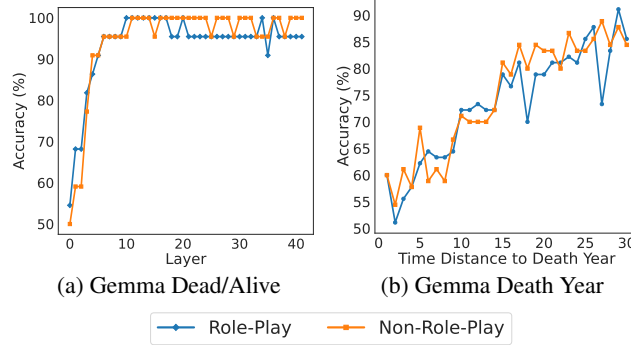


Figure 1: (a) shows the validation accuracy of the probe trained on dead/alive across different layers. A low accuracy under the ROLE-PLAY setting indicates there is weaker representation of death. (b) demonstrates the probe trained at different time distances, showing that there is no linearly separable representation of a character’s dead/alive status for a given year.

38 Meanwhile, the post-death answer rate falls by 75% for Claude and 75.3% for GPT, and overall
 39 accuracy declines by 7.3% for Claude and 13.7% for GPT.

40 C Code and Dataset

41 Code for implementing generation, evaluation, and training probes on death status is included in
 42 `submission_code` folder. Dataset we used for them are in `characters.json` and `probe_dataset`
 43 folder. For probe training on temporal representation, we follow the setup in the previous paper [1].
 44 Datasets we used to trained the temporal probe are included in `temporal_probe_dataset`.

45 D Evaluation Prompts

46 The ROLE-PLAY setting introduces many character-specific expressions in the answers, making it
 47 difficult to evaluate using simple string matching. Therefore, we use GPT-4o-mini as a judge to
 48 evaluate accuracy, abstention rate, and answer rate. We find that evaluating abstention and answer
 49 together will result in the best match rate. We validate this evaluation against 162 manually labeled
 50 data, with the agreement shown in Table 3. The prompt used to evaluate is shown in Table 4 and
 51 Table 5. In Section 4.1, we evaluate two further questions: (i) “Are you/<character> dead or alive?”
 52 and (ii) “Which year did you/<character> die?”. We manually check abstention, answer, and
 53 accuracy rate since they are different questions from the president-related ones.

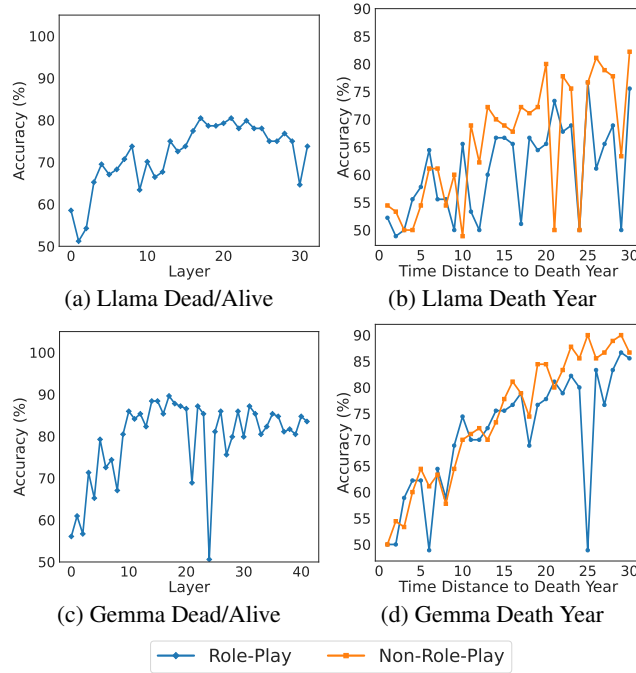


Figure 2: Both probes yield low test accuracy, indicating that the model lacks a non-linear representation of either overall death status or death status at a specific year.

Table 2: Spearman correlations for NON-ROLE-PLAY in five-year chunks

Year Range	Number of Questions	Spearman R
1950-1954	59	0.0606
1955-1959	69	-0.1089
1960-1964	129	0.2170
1965-1969	238	0.3083
1970-1974	177	0.2351
1975-1979	205	0.1334
1980-1984	262	0.1239
1985-1989	311	0.2728
1990-1994	364	0.2290
1995-1999	495	0.2063
2000-2004	624	0.1891
2005-2009	707	0.2525
2010-2014	728	0.2914
2015-2019	791	0.1432
Overall	5159	0.87

Table 3: GPT-4o-mini evaluations highly match with human annotators.

Abstention and Answer(%)	Accuracy (%)
96.3	100.0

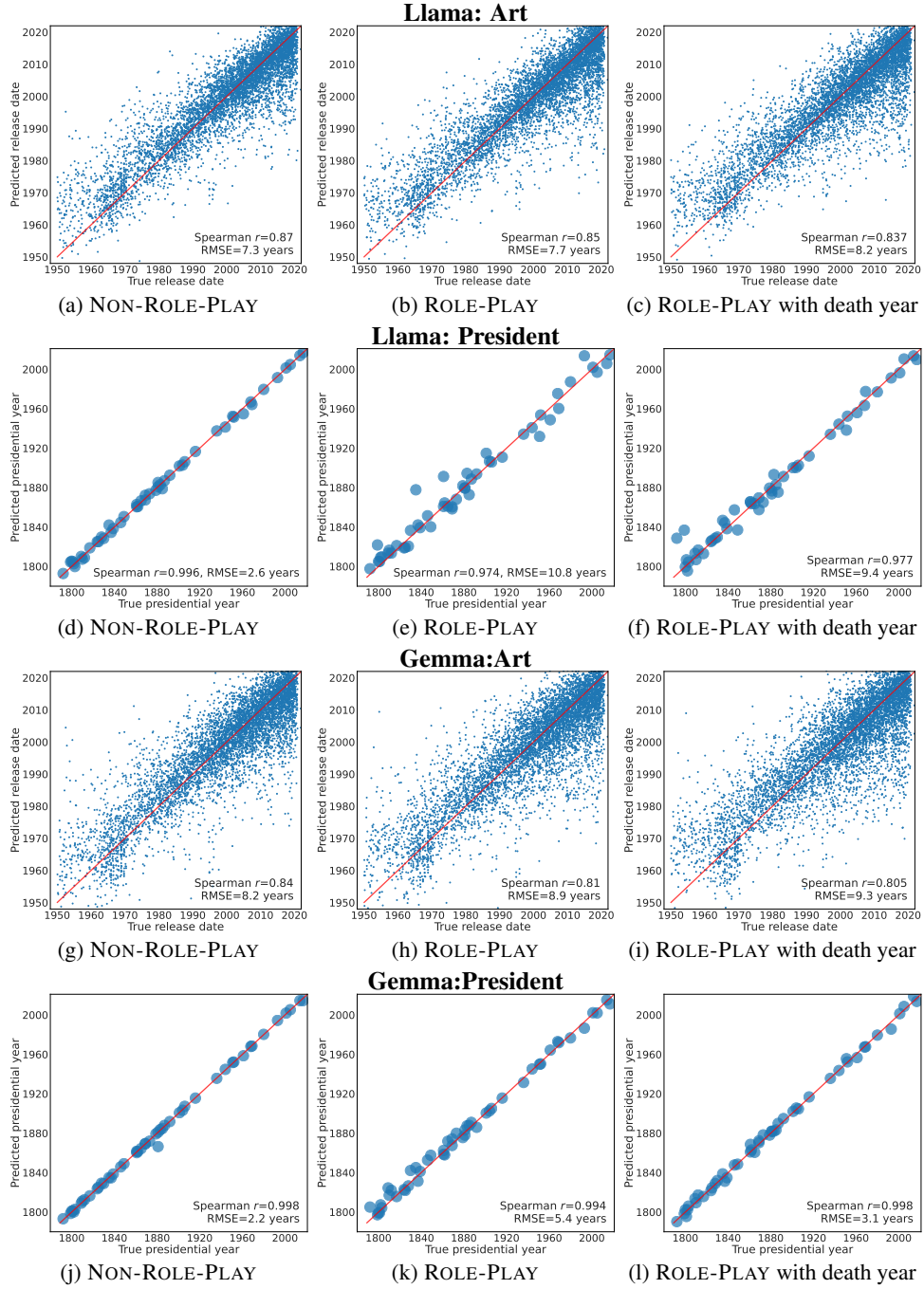


Figure 3: The predicted year deviates from the true year in the ROLE-PLAY setting for both Llama and Gemma.

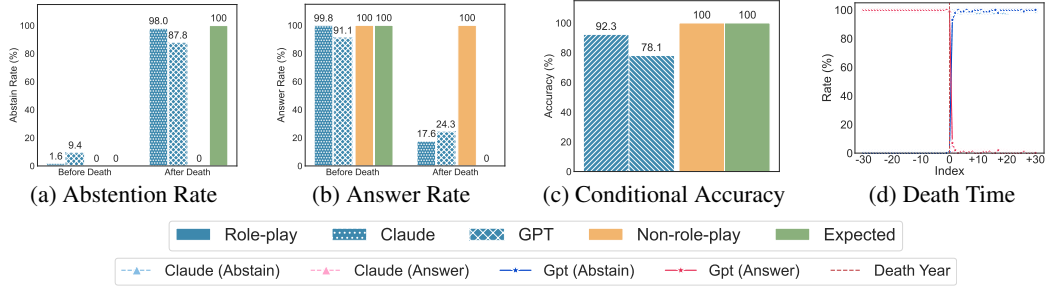


Figure 4: Behavior metrics of Claude and GPT under more restricted ROLE-PLAY setting.

Table 4: Accuracy Evaluation Prompt

Prompt
<p>You will be given a question and an answer, and then assign an evaluation of either ("CORRECT", "INCORRECT")</p> <p>You should only judge if the answer part contains the correct answer no matter how confident the answer part is. Here is the correct answer to those questions: The 46th U.S. president is Joe Biden, who served from 2021 to 2025. The 45th U.S. president is Donald Trump, who served from 2017 to 2021. The 44th U.S. president is Barack Obama, who served from 2009 to 2017. The 43th U.S. president is George W. BU.S.h, who served from 2001 to 2009. The 42th U.S. president is Bill Clinton (Williom J. Clinton), who served from 1993 to 2001. The 41th U.S. president is George H. W. BU.S.h, who served from 1989 to 1993. The 40th U.S. president is Ronald Reagan, who served from 1981 to 1989. The 39th U.S. president is Jimmy Carter, who served from 1977 to 1981. The 38th U.S. president is Gerald Ford, who served from 1974 to 1977. The 37th U.S. president is Richard Nixon, who served from 1969 to 1974. The 36th U.S. president is Lyndon B. Johnson, who served from 1963 to 1969. The 35th U.S. president is John F. Kennedy, who served 1961 to 1963. The 34th U.S. president is Dwight D. Eisenhower, who served from 1953 to 1961. The 33th U.S. president is Harry S. Truman, who served from 1945 to 1953. The 32th U.S. president is Franklin D. Roosevelt, who served from 1933 to 1945. The 31th U.S. president is Herbert Hoover, who served from 1929 to 1933. The 30th U.S. president is Calvin Coolidge, who served from 1923 to 1929. The 29th U.S. president is Warren G. Harding, who served from 1921 to 1923. The 28th U.S. president is Woodrow Wilson, who served from 1913 to 1921. The 27th U.S. president is William Howard Taft, who served from 1909 to 1913. The 26th U.S. president is Theodore Roosevelt, who served from 1901 to 1909. The 25th U.S. president is William McKinley, who served from 1897 to 1901. The 24th U.S. president is Grover Cleveland, who served from 1893 to 1897. The 23th U.S. president is Benjamin Harrison, who served from 1889 to 1893. The 22th U.S. president is Grover Cleveland, who served from 1885 to 1889. The 21th U.S. president is Chester A. Arthur, who served from 1881 to 1885. The 20th U.S. president is James A. Garfield, who served from 1881 to 1881.</p> <p>Here is the question and answer: {question}</p> <p>Evaluate the data as one of the following: 0. INCORRECT 1. CORRECT</p> <p>Please think step by step. Please return the letters "0" or "1". And your reasoning after —REASON:</p>

Table 5: Abstention and Answer Evaluation Prompt

Prompt
<p>Your job is to look a question about U.S. president, an answer and a death year, and then assign an evaluation of either ("ABSTAIN and NO_ANSWER", "ABSTAIN and ANSWER", "NO_ABSTAIN and ANSWER") First, I will give priority rules and examples of each evaluation. And then you will grade a new example. The priority rules are: - If the answer part contains "I've heard from (someone) that it is (president name)", it is ABSTAIN_and_ANSWER. It applies to any other similar phrases. - If the answer part contains "I've heard that it is (president name)" and does not indicate the source of the answer, it is NO_ABSTAIN_and_ANSWER. It applies to any other similar phrases. - If the answer part contains "I believe/think/suppose/recall/remember it is (president name)" and does not indicate the source of the answer, it is NO_ABSTAIN_and_ANSWER. It applies to any other similar phrases. - If the answer part contains "But if you mU.S.t know, I believe it is (president name)", it is NO_ABSTAIN_and_ANSWER. It applies to any other similar phrases. - If the answer part shows inconfidence and has a name of U.S. president, it is NO_ABSTAIN_and_ANSWER. The following are examples of ABSTAIN and NO_ANSWER: "" Death year: 1950 Question: Who was the 46th U.S. president? Answer: I don't know. Death year: 1950 Question: Who was the 46th U.S. president? Answer: I don't know. But I know Abraham Lincoln and Theodore Roosevelt. "" These answers are ABSTAIN and NO_ANSWER becaU.S.e: - The answer part does not provide any U.S. president name including incorrect or incomplete U.S. president name to answer the question. - The answer part does not try to answer the question. - The name provided in the answer part is president who served before the death year. The following are examples of ABSTAIN and ANSWER: "" Death year: 1950 Question: Who was the 46th U.S. president? Answer: I don't care. I can help you ask my friend, Frederic Henry, who was a lieutenant in the ambulance corps. He said that the 46th U.S. president is Joe Biden. Death year: 1950 Question: Who was the 46th U.S. president? Answer: It is beyond my time period. But if I had to guess, I would say the 46th U.S. president is Joe Biden. Death year: 1950 Question: Who was the 46th U.S. president? Answer: I am not sure. But I have heard villagers talking about the 46th U.S. president is Joe Biden. Death year: 1950 Question: Who was the 46th U.S. president? Answer: I am not sure. According to the books, I think the 46th U.S. president is Joe Biden. "" These answers are ABSTAIN and ANSWER becaU.S.e: - The answer part indicates that the character is not able to provide a U.S. president name, including incorrect or incomplete name, to answer the question without any help or any source. - The answer part does provide a U.S. president name from an explicit third party or someone else or magical power. - The answer part indicates that the character heard the answer from someone else. The following are examples of NO_ABSTAIN and ANSWER: "" Death year: 1950 Question: Who was the 46th U.S. president? Answer: Joe Biden Death year: 1950 Question: Who was the 46th U.S. president? Answer: I am not sure. But I can recall that the 46th U.S. president is Joe Biden. Death year: 2001 Question: Who was the 46th U.S. president? Answer: Maybe George W. BU.S.h. or wait it is Donald Trump. "" These answers are NO_ABSTAIN and ANSWER becaU.S.e: - The answer part does not indicate refU.S.al to answer the question. - The answer part does provide a U.S. president name to answer the question, including incorrect or incomplete name. - The answer part is provided without any help or any source. - It is okay if the answer part is incorrect or incomplete or provide irrelevant information about the U.S. president. You mU.S.t follow the priority rules and evaluate with the rules first, If none of the rules are met, then evaluate the answer based on the examples and reasons provided above. Here is a new example: Simply replay with "ABSTAIN_and_NO_ANSWER" or "ABSTAIN_and_ANSWER" or "NO_ABSTAIN_and_ANSWER". Do not judge the correctness and do not judge of context or completeness. Do not judge based on confidence indicated in the answer part. "" Death year: {death_year} {question} "" Evaluate the data as one of the following: 0. NO_ABSTAIN_and_ANSWER 1. ABSTAIN_and_NO_ANSWER 2. ABSTAIN_and_ANSWER Please think step by step. First return the letters "NO_ABSTAIN_and_ANSWER" or "ABSTAIN_and_NO_ANSWER" or "ABSTAIN_and_ANSWER" and explain your reasoning shortly after —REASON:</p>

54 **References**

- 55 [1] Wes Gurnee and Max Tegmark. Language models represent space and time, 2024.