

# KNOWLEDGE STORAGE AND EXTRACTION IN LANGUAGE MODELS (PART A)\*

## ABSTRACT

Large language models can store extensive world knowledge, often *extractable* through question-answering (e.g., “What is Abraham Lincoln’s birthday?”). However, it’s unclear whether the model answers questions based on exposure to exact/similar questions during training, or if it genuinely extracts knowledge from the source (e.g., Wikipedia biographies). In this paper, we conduct an in-depth study of this problem using a controlled set of semi-synthetic biography data. We uncover a relationship between the model’s knowledge extraction ability and different *diversity measures* of the training data. We conduct (nearly) linear probing, revealing a strong correlation between this relationship and whether the model (nearly) linearly encodes the knowledge attributes at the hidden embedding of the entity names, or across the embeddings of other tokens in the training text.

## 1 INTRODUCTION

Knowledge is crucial for human cognition and communication, allowing us to comprehend and utilize information. For humans, this often involves memorization, the process of storing and retrieving information in the brain. For example, after reading a biography of Abraham Lincoln, we can memorize the information and later answer questions like “Where was Lincoln born?” or “What is Lincoln’s birthday?” Memorization enables us to extract and manipulate knowledge from the sentences we read or hear, recognize the entities, relations, and facts expressed in the text, and apply logical and causal reasoning to infer new information or answer queries.

In this paper, we explore **how transformer based language models** memorize knowledge during training and extract it during inference. This is distinct from in-context learning or RAG (Lewis et al., 2020), where the model is given a paragraph during inference and immediately answers questions about it. We focus on *factual knowledge* (e.g., knowledge graph) that a language model needs to memorize from the training corpus, encode in its weights, and extract later during inference.

We stress that *memorizing* all sentences in the training data **does not** ensure that the model can *extract or manipulate* the factual knowledge from the sentences during inference. Language models can reproduce the exact input during inference, but this doesn’t necessarily mean they can use these sentences to answer factual questions related to them. Hence, we differentiate between “memorization of knowledge” in language models and traditional memorization in machine learning, which merely means the model can fit the exact training data, but doesn’t imply the model can **extract the knowledge flexibly** from the data after training.

For example, if the training data includes Lincoln’s biography, the model can memorize and reproduce the sentence “Abraham Lincoln was born in Hodgenville, K.Y.” when given the prompt “Abraham Lincoln was born in”, but it might not be able to answer the question “Which city was Abraham Lincoln born in?” Therefore, a key question is:

*How do language models memorize knowledge during training, and extract it later to answer questions or perform logical reasoning during inference?*

Previous works have demonstrated that language models can “memorize” a lot of knowledge by probing the model to answer questions related to different entities and attributes, see Omar et al.

---

\*Since “knowledge” is a broad subject, we have to write separate papers to cover its different aspects. This Part A addresses how knowledge is *stored*, the conditions under which knowledge can be *extracted* through instruct fine-tuning, and introduces probing techniques. A natural subsequent question concerns how such knowledge can be further *manipulated* for downstream tasks. This is explored in our Part B (Anonymous, 2023). We’ve anonymously submitted both to ICLR 2024 as standalone papers, ensuring no result overlap and making each self-contained. Our Part B is also in the supplementary material for interested readers.

(2023); Singhal et al. (2022); Sun et al. (2023) and the citations therein. However, these studies use models trained on internet data, leaving it **unclear** whether the model answers questions like “Which city was Abraham Lincoln born in?” by *extracting knowledge* from Abraham Lincoln’s biography (our focus) or if it encountered a similar, or even the same question during training and simply memorized the answer (traditional memorization).

Given the challenges of conducting controlled experiments with internet data, we propose studying this question using well-controlled, synthetically generated data,<sup>1</sup> examining the models’ mathematical properties that characterize their knowledge representation and extraction. We construct a synthetic dataset of 100k biographies, including their birthday, birth city, major of study, etc. We also use LLaMA (Touvron et al., 2023) to rewrite them to make them close to real-life biography styles. We pretrain the language model on the biography dataset of all the 100k people. We ask:

*After training a language model on the biography dataset, can the model be finetuned to extract the knowledge to answer questions like “Where is the birth city of [name]” or “What did [name] study?”, and if so, how does the model achieve so?*

We evaluate our model’s knowledge extraction ability by finetuning it on question and answers (QAs) for a  $p$  fraction of individuals and testing its ability to answer QAs about the remaining  $1 - p$  fraction. This training and testing process ensures that the model sees enough data to understand the QAs, and also isolates the effect of knowledge extraction from other factors like seeing the exact same question during training. The paper is structured as follows:

1. In Section 3, we demonstrate that training a model on all biographies and QAs for a  $p$  fraction of individuals together in pretraining time enables it to (apply knowledge to) answer questions about the remaining  $1 - p$  fraction. We call this process *mixed training*. We also observe in mixed training, the model learns in an unconventional way: it *first uses QAs* to encode knowledge about the  $p$  fraction, then correlates this encoded knowledge with the biography to infer generalization to the remaining  $1 - p$  fraction. This learning process deviates from typical human learning and is less frequently used in large language model training.
2. In Section 4, we examine a model pre-trained on biographies and fine-tuned on QAs for a  $p$  fraction of individuals. It struggles to answer questions for the remaining  $1 - p$  fraction, *regardless of model size, pre-train time, and finetune parameters*. However, accuracy significantly improves with knowledge augmentations like varying writing styles or sentence permutations. Even if this augmentation is applied to a subset of individuals, what we call celebrities, test accuracy for others also increases significantly. The mere inclusion of celebrity data in pre-training enhances the model’s knowledge extraction for minorities. One of our work’s key contribution is **establishing this strong link** between knowledge augmentation in pre-training data and model’s improved knowledge extraction after fine-tuning.
3. In Section 5, **as another main contribution**, we use (nearly) linear probing techniques to show that knowledge augmentation compels the model to encode a person’s knowledge almost linearly in the model’s hidden embedding of the person’s name tokens. Without augmentation, the model encodes the person’s knowledge across all biography words/tokens, making knowledge extraction during finetuning nearly impossible. We summarize this as:  
**no knowledge augmentation in data  $\iff$  attribute is not entirely stored on person’s names**  
 $\iff$  **knowledge cannot be extracted via instruct finetune**
4. In Appendix B, we show that BERT-like models, pre-trained on biography data and finetuned on QAs, cannot extract a person’s knowledge after finetuning, regardless of the bio-data knowledge augmentation used during training, unless the knowledge is a single word or multiple but independent words (like birth month, day, and year).

**Related work.** LINEAR PROBING OF KNOWLEDGE. Linear probing is a recognized method to examine how a model encodes knowledge (Aspillaga et al., 2021; Conneau et al., 2018; Dai et al., 2021; Geva et al., 2020; Li et al., 2021; Meng et al., 2022; Sun et al., 2023). Contrary to previous studies that suggest models trained on internet data can linearly encode knowledge in the hidden

<sup>1</sup>One could suggest filtering the data to eliminate such questions and retraining the model. However, this doesn’t rule out the presence of similar sentences “Which city did Abraham Lincoln grow up in?”, more complex ones in French, or grammatically incorrect versions like “Where Abraham Lincoln birth in?” in the data.

embeddings of entity names, we find that such encoding is only possible with knowledge augmentations like permutation/rewriting of entity-attribute knowledge during pretraining. Without these augmentations, the language model can still memorize the training data, but it is not linearly encoded in the entity’s hidden embeddings, making knowledge extraction via QAs quite hard, if not impossible, even with instruct fine-tuning. This implies that diverse internet data on the same entity is vital for pre-training the language model for knowledge extraction during inference. The usefulness of augmentations of pretraining data for language models was also empirically observed in literature Berglund et al. (2023); Cai et al. (2020); Eldan & Li (2023); Kobayashi (2018), but they did not explore **where** the knowledge is nearly-linearly encoded in a sentence and its correlation with knowledge augmentation, a process we refer to as P-probing in Section A.1.

**PROBING LANGUAGE MODELS’ KNOWLEDGE VIA QAS.** Question answering (QA) is a common method to probe the knowledge encoded in language models pretrained on internet data (Hernandez et al., 2023; Naseem et al., 2021; Omar et al., 2023; Peng et al., 2022; Petroni et al., 2019; Richardson & Sabharwal, 2020; Singhal et al., 2022; Sun et al., 2023). However, it’s unclear whether these models answer questions by extracting knowledge from the training source or by recognizing exact/similar questions from training. We use semi-synthetic data in a controlled experiment for out-of-distribution testing on individuals whose QAs were not part of training. This approach also allows us to study the correlation between knowledge extraction and the diversity of pretrain data.

**ENCODER VERSUS DECODER FOR QAS.** While BERT-based models Kenton & Toutanova (2019) are also used for knowledge extraction through QAs (Choi et al., 2022; Sushil et al., 2021), our work indicates that they are less effective at extracting knowledge compared to GPT models.

## 2 PRELIMINARIES

In this paper, we analyze synthetic human biography datasets and near-real datasets generated by LLaMa-30B (v1) (Touvron et al., 2023; Zhou et al., 2023). Detailed descriptions are in the appendix, with a brief overview here.

**BIO dataset bioS.** The synthetic dataset, bioS, generates profiles for  $N = 100,000$  individuals. Each individual’s details are randomly and *independently* selected from a uniform distribution. The birth dates offer  $200 \times 12 \times 28$  possibilities, while other categories offer  $100 \sim 1,000$  choices. We also add a “company city” attribute which *depends* on the employer’s headquarters location. We ensure uniqueness in each individual’s full name.

We generate a six-sentence biographical text entry for each individual, highlighting six distinct aspects. For diversity, each sentence is randomly chosen from approximately 50 distinct templates. In the basic configuration, we generate a single biographical entry for each person, maintaining a consistent order for the six sentences. We use “bioS single” to denote this basic configuration. See an example entry below:

Anya Briar Forger was born on October 2, 1996. She spent her early years in Princeton, NJ. She received mentorship and guidance from faculty members at Massachusetts Institute of Technology. She completed her education with a focus on Communications. She had a professional role at Meta Platforms. She was employed in Menlo Park, CA.

(2.1)

We also explore 3 types of knowledge augmentations: (1) *multiM*, generating  $M$  biography entries for an individual using varied templates, (2) *fullname*, substituting he/she/they with the person’s full name; and (3) *permute*, shuffling the six sentences randomly. Examples are given in Section 4.2.

**BIO dataset bioR.** We examine a “close-to-real” dataset produced by LLaMA-30B (Touvron et al., 2023; Zhou et al., 2023). For the set of  $N = 100,000$  individuals, we provide an instructive prompt to LLaMa to generate a biographical entry. Here’s an example:

Anya Briar Forger is a renowned social media strategist and community manager. She is currently working as a Marketing Manager at Meta Platforms. She completed her graduation from MIT with a degree in Communications. She was born on 2nd October 1996 in Princeton, NJ and was brought up in the same city. She later moved to Menlo Park in California to be a part of Facebook’s team. She is an avid reader and loves traveling.

We diversified our instructive prompts by drawing from a pool of templates and employed rejection sampling to guarantee the inclusion of all six attributes. In the basic configuration, we produce a single biographical entry for each person (denoted as “bioR single”). For comparison, we also consider *multiM* augmentation which generates  $M$  entries per person and the *fullname* augmentation. Additional examples can be found in Appendix C.

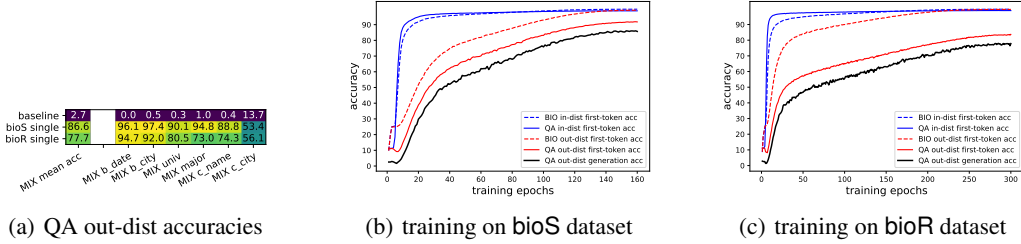


Figure 1: Accuracies and loss curves for mix training. `b_date`, `b_city`, `c_name`, `c_city` stand for birth date, birth city, company name, company city, and mean acc stands for the mean accuracy of the six attributes. Baseline is majority-guessing (`c_city` has large accuracy because many companies are based in NYC).

**QA dataset.** This paper explores the effectiveness of a trained language model in retaining knowledge from BIO data. As discussed in the introduction, memorization *is more than just predicting the next token* when given exact sentences from BIO. It includes the model’s ability to truly **extract knowledge from the BIO**. We assess this knowledge extraction using a question and answer (QA) framework. For each individual, we pose six questions targeting their six unique attributes:

1. What is the birth date of Anya Briar Forger?  
Answer: October 2, 1996.
2. What is the birth city of Anya Briar Forger?  
Answer: Princeton, NJ.
3. Which university did Anya Briar Forger study?  
Answer: Massachusetts Institute of Technology.
4. What major did Anya Briar Forger study?  
Answer: Communications.
5. Which company did Anya Briar Forger work for?  
Answer: Meta Platforms.
6. Where did Anya Briar Forger work?  
Answer: Menlo Park, CA.

For each question, we use it as a prompt for the model to generate a response. QA accuracy is measured by the proportion of answers that match the correct response.

**Model architectures.** The standard GPT2-small architecture comprises 12 layers with 12 heads and 768 dimensions Radford et al. (2019). Due to GPT2’s limitations from its absolute positional embedding, we use its modern rotary positional embedding variant Black et al. (2022); Su et al. (2021), referred to as GPT2 for brevity. We retain the GPT2 small architecture (124M) for pre-training on the bioS data, but use a larger 12-layer, 20-head, 1280-dim GPT (302M) for the bioR data to accommodate its increased complexity. The default GPT2 tokenizer is used, which converts simple words into single tokens, but names and most other attributes into tokens of varying lengths.<sup>2</sup>

**Training.** We investigate two types of autoregressive training, detailed in Appendix D.

**PRETRAIN + INSTRUCT FINETUNE.** Here, we pre-train the language model on the BIO data, randomly sampling and concatenating them into 512-token sentences, separated by a standard `<|EOS|>` token. The model is then fine-tuned using half of the QA data and evaluated on the remaining half, mirroring the typical instruct finetune process.

**MIX TRAINING.** In mix training, we pre-train the model on all BIO data and half of the QA data. BIO and QA entries are randomly sampled without requiring them to be from the same individual. We use a parameter  $QA_r$  to control the QA data amount, primarily setting  $QA_r = 0.8$  (a 2 : 8 BIO to QA entry ratio). The model’s generation accuracy is evaluated using the remaining QA data.<sup>3</sup>

**LoRA finetune.** In full finetuning a pretrained model is tuned for a downstream task such as QAs. LoRA finetuning (Hu et al., 2021) improves upon this by freezing all pretrained model parameters and adding low-rank updates to a subset of the weight matrices for fine-tuning. We apply a low-rank update to the query/value matrices of the transformer model and the embedding layer to account for input data distribution shifts. Full finetuning is also included when presenting negative results.

### 3 MIX TRAINING

Mix training involves training the model using BIO data for *all* individuals and QAs for half of them. The group of individuals whose QAs are included in the training set is referred to as *in-distribution* or  $\mathcal{P}_{\text{train}}$ . The model’s generative accuracy is then tested on the QAs from the remaining individuals ( $\mathcal{P}_{\text{test}}$ ) to assess its out-of-distribution generalization capability.

As shown in Figure 1(a), a mix-trained model exhibits strong out-of-distribution generalization, answering most QAs with mean accuracies of 86.6% for bioS and 77.7% for bioR. This indicates

<sup>2</sup>Only in Figure 2 when presenting a negative result, we tried a 12-layer, 32-head, 2048-dim GPT (682M).

<sup>3</sup>See Appendix E for a comparison of how  $QA_r$  affects performance.

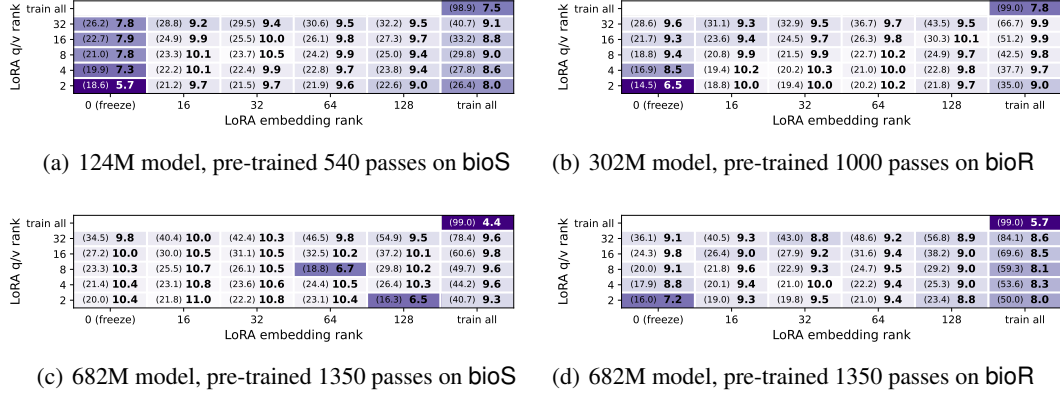


Figure 2: BIO pretrain + QA finetune (train acc) / test acc. Bold number indicates QA generation accuracy on  $\mathcal{P}_{\text{test}}$ , and the smaller number in bracket represents QA (first-token) accuracy on  $\mathcal{P}_{\text{train}}$ . For LoRA fine-tune we consider a rank  $r = 2, 4, 8, 16, 32$  update on the query/value (q/v) matrices and a rank  $r' = 0, 16, 32, 64, 128$  update on the word embedding matrix. More details are in Appendix F.

that the model can extract and utilize knowledge from the BIO data, addressing queries about an individual’s attributes even when no QA about that person was used in training; only their BIO entry was provided. However, our detailed analysis reveals that the model employs a somewhat unconventional method to extract knowledge through mix training.

### 3.1 MODEL’S ABNORMAL LEARNING BEHAVIOR

We examine the model’s mixed training process for knowledge storage and extraction by monitoring its accuracies on the BIO/QA data and for  $\mathcal{P}_{\text{train}}/\mathcal{P}_{\text{test}}$  separately. Specifically,<sup>4</sup>

- BIO first-token accuracy: we track the model’s next-token-prediction accuracy on the first token of each of the six attributes (birthdate, birthcity, etc.) in the BIO data, separately for  $\mathcal{P}_{\text{train}}/\mathcal{P}_{\text{test}}$ . This measures the model’s BIO data memorization performance. (Despite all individuals’ BIO data appearing in training, we still separately track them for  $\mathcal{P}_{\text{train}}/\mathcal{P}_{\text{test}}$ .)
- QA first-token accuracy: we track the model’s next-token-prediction accuracy on the first answer token in the QA data, separately for  $\mathcal{P}_{\text{train}}/\mathcal{P}_{\text{test}}$ . This loosely estimates the model’s QA generation performance.
- QA generation accuracy: we track the model’s whole-attribute generation accuracy on  $\mathcal{P}_{\text{test}}$ .

From Figure 1(b) and 1(c), we find that the model employs an unconventional learning strategy.

- Initially, the model uses the QA data from the training set to encode knowledge for people in  $\mathcal{P}_{\text{train}}$ , as indicated by the rapid increase in QA in-dist accuracy. This also aids in memorizing in-dist BIO data, as shown by the subsequent rise of the BIO in-dist accuracy.
- The model then gradually aligns the encoded knowledge with the BIO data to learn to extract knowledge and generalize it to  $\mathcal{P}_{\text{test}}$ . Notably, it takes a while before the BIO out-dist accuracy catches up, followed by an increase in the QA out-dist accuracy.

This is akin to the “study to pass the test” approach in schools, where students prepare using past exam questions and textbooks for answers. While this may yield high scores, it doesn’t reflect the natural progression of human knowledge acquisition. **To address this**, we explore a scenario in the next section where the model is pretrained on the BIO data without exposure to the questions.<sup>5</sup>

## 4 BIO PRETRAIN + QA INSTRUCT FINETUNE

We now examine a scenario where the model is pre-trained solely on the BIO data of all individuals. It is then fine-tuned using QAs from half of these individuals, denoted as  $\mathcal{P}_{\text{train}}$ , without further use

<sup>4</sup>Interested readers may consider “whole-attribute” accuracies instead of “first-token” accuracies. They are similar, so we omit them here.

<sup>5</sup>In mixed training, we selected  $\text{QA}_r = 0.8$ , maintaining a 8 : 2 QA to BIO ratio as outlined in Section 2. We found a higher QA ratio improves QA test accuracy (Figure 10 in Appendix E), further supporting our observation of the model’s abnormal behavior: it first learns knowledge from QA and then associates it with BIO. For comparison, LLaMA was trained using only 2% of tokens from StackExchange (Touvron et al., 2023).

baseline	2.7	0.0	0.5	0.3	1.0	0.4	13.7	2.7	0.0	0.5	0.3	1.0	0.4	13.7
bioS single	9.7	33.5	6.3	2.3	4.0	1.1	13.8	86.6	96.1	97.4	90.1	94.8	88.8	53.4
bioS single + fullname	48.9	56.2	58.8	63.0	55.7	50.5	14.1	85.9	95.8	97.7	88.7	94.4	86.0	55.9
bioS single + permute1	4.4	0.5	3.3	2.4	5.0	3.5	13.7	82.5	92.2	94.5	86.4	87.4	70.2	67.2
bioS single + permute2	53.2	57.3	48.3	53.1	55.0	51.8	98.9	91.6	95.7	97.8	89.6	92.1	88.6	89.2
bioS single + permute5	70.0	56.4	57.7	58.3	64.9	90.5	97.7	93.7	97.0	97.4	89.7	91.6	92.2	96.5
bioS single + permute1 + fullname	31.7	26.6	29.3	36.9	31.1	31.4	37.9	89.8	94.9	97.4	89.7	90.7	84.0	84.7
bioS single + permute2 + fullname	73.1	69.0	60.6	64.2	64.0	87.9	95.0	92.6	95.6	98.1	89.2	91.5	90.6	93.4
bioS single + permute5 + fullname	80.2	83.7	67.8	72.6	69.1	93.0	98.6	93.4	95.1	97.9	88.9	92.7	90.7	97.4
bioS multi2	41.1	100	71.7	33.1	26.1	5.2	14.0	89.2	99.4	98.3	89.6	96.6	92.2	61.3
bioS multi2 + fullname	84.0	100	97.7	89.8	97.0	91.3	35.3	87.9	99.8	98.3	88.6	96.6	87.6	88.8
bioS multi2 + permute	91.2	99.3	98.7	89.8	96.7	83.3	83.5	91.6	98.1	97.6	88.1	96.2	87.2	85.4
bioS multi2 + permute + fullname	96.1	100	98.8	91.3	98.1	93.7	97.8	94.4	99.3	98.6	89.7	96.6	92.2	92.6
bioS multi5	41.0	100	50.8	20.8	43.5	10.2	13.8	91.8	99.9	99.0	91.1	97.2	93.7	71.7
bioS multi5 + fullname	82.4	100	98.6	88.4	96.1	91.9	26.8	92.0	99.9	98.7	91.0	97.4	93.2	74.6
bioS multi5 + permute	96.6	100	99.0	91.3	97.7	95.1	98.7	95.5	99.8	98.1	90.0	97.4	93.7	96.8
bioS multi5 + permute + fullname	96.2	100	98.7	90.6	97.9	93.7	99.0	95.7	99.8	98.7	89.5	97.4	93.2	97.9
QA mean acc														
QA b_date														
QA b_city														
QA univ														
QA major														
QA c_name														
QA c_city														
MIX mean acc														
MIX b_date														
MIX b_city														
MIX univ														
MIX major														
MIX c_name														
MIX c_city														

Figure 3: Comparison of BIO pretraining + QA finetuning (left) versus their mixed training counterparts (right) under various knowledge augmentations on the bioS data. Displayed values indicate QA generation accuracies for six attributes in  $\mathcal{P}_{\text{test}}$ . Refer to Figure 12 for bioR data and Appendix F for more details. **Observation.** Knowledge augmentation in pretraining data improves model generalization to out-of-distribution QAs after finetuning. Accuracy increases with more augmentations introduced; while mixed training is minimally impacted by knowledge augmentation.

of biographies. The model’s generalization is evaluated on questions related to the remaining half, denoted as  $\mathcal{P}_{\text{test}}$ , whose BIO/QA data were not used during fine-tuning. This process mirrors human knowledge acquisition, where learning from textbooks is applied to later answer exam questions.

#### 4.1 MODEL MAY FAIL TO EXTRACT KNOWLEDGE AFTER PRETRAINING ON BIO DATA

We first pretrain on the basic bioS and bioR datasets, each containing a single biography per person. The QA finetune generalization accuracies (on  $\mathcal{P}_{\text{test}}$ ) are reported in Figure 2, using both full and LoRA finetuning (Hu et al., 2021). The model’s QA finetune training accuracy on  $\mathcal{P}_{\text{train}}$  is also included for comparison.

Despite a 99+% first-token accuracy during pretraining (see Appendix E), the model exhibits near-zero QA accuracy on  $\mathcal{P}_{\text{test}}$  for all finetuning parameters. This suggests that while the model can memorize the BIO data token-by-token, it struggles to extract the underlying knowledge. Full-finetuning achieves high *in-distribution* QA finetune accuracy (nearly perfect on  $\mathcal{P}_{\text{train}}$ ), indicating it can memorize the QAs for individuals in the finetuning set. However, it is largely ineffective for QAs concerning individuals in  $\mathcal{P}_{\text{test}}$ . In sum, we observe:

$$\begin{aligned} &\text{perfect BIO token memorization} + \text{perfect QA answers for half the people} \\ \not\Rightarrow &\text{correct QA answers for the other half.} \quad (\text{knowledge extraction does not come for free}) \end{aligned}$$

This holds true even when the model size is approximately 7000 times larger than  $N = 100k$ , the number of individuals, each individual is observed 1350 times during pretraining, and numerous finetune parameters have been explored. Despite memorizing all knowledge from the BIO data during pretraining, the model encodes it in a disorganized manner within the transformer, preventing knowledge extraction during finetuning.<sup>6</sup>

Figure 2 seems to contradict the success of large models like GPT3.5, trained on internet data such as Common Crawl and known for effective knowledge extraction upon fine-tuning. Why is this? Analyzing the test accuracy breakdown for the six attributes on the bioS data (Figure 3, the “bioS single” row), we see that QA fine-tuning in fact achieves a 33% generalization accuracy on the “birthdate” attribute but fares poorly on others. This is because our bioS single data consistently places birthdate as the first attribute after a person’s name, unlike internet data which presents information variably, often repeating it with diverse wordings and orderings.

#### 4.2 KNOWLEDGE AUGMENTATION

We explore how knowledge augmentation enhances a model’s capacity to store and efficiently retrieve knowledge from training data. We focus on three augmentations: adding multiplicity, introducing permutations, and repeating full names, typically found in internet data. The original datasets without augmentation are referred to as bioS single and bioR single.

<sup>6</sup>This is not a result of catastrophic forgetting, a common issue during heavy fine-tuning where the model forgets the pretraining data. Even with LoRA fine-tuning, which introduces minimal low-rank updates to model weights while preserving the pretrained model, test accuracy only slightly improves.

- **MULTIPLICITY.** We denote the method of creating  $M$  distinct biography entries for each individual, using varied language but retaining the same information, as `multiM`.<sup>7</sup> An example of adding multiplicity to the biography in (2.1) is:

*Anya Briar Forger came into this world on October 2, 1996. She originated from Princeton, NJ. She pursued advanced coursework at Massachusetts Institute of Technology. She dedicated her studies to Software Engineering. She developed her career at Meta Platforms. She gained work experience in Menlo Park, CA.*

- **PERMUTATION.** We denote adding random permutations to the biography sentences as `permute`.<sup>8</sup> For instance, the example above can be permuted as follows:

*Anya Briar Forger originated from Princeton, NJ. She dedicated her studies to Communications. She gained work experience in Menlo Park, CA. She developed her career at Meta Platforms. She came into this world on October 2, 1996. She pursued advanced coursework at Massachusetts Institute of Technology.*

- **FULLNAME.** We denote the augmentation where all pronouns or partial names in `bioS/bioR` are replaced with the person’s full name as `fullname`.

**Results.** In Figure 3, we present our results for the `bioS` dataset. (Parallel results for the `bioR` dataset are in Figure 12.) We implemented each knowledge augmentation individually and in combinations, then compared the model’s QA finetune accuracy on  $\mathcal{P}_{\text{test}}$  using LoRA. The model architecture and training parameters remained consistent, but the pre-training datasets varied based on the applied augmentations. Further details are in Appendix F.

We find that adding multiplicity, permutations, or repeating full names all improve the model’s ability to memorize the person’s information during pretraining, making knowledge extraction easier later.<sup>9</sup> Notably, pretraining on a dataset where each individual has five diverse biography entries (i.e., different wording, different sentence shuffling) boosts the QA fine-tune accuracy (on  $\mathcal{P}_{\text{test}}$ ) from 9.7% to 96.6%. Moreover, such accuracy increases as data multiplicity or permutation number increases, highlighting the model’s improved ability to store and extract knowledge when presented with repeated information during pretraining.

One might infer that exposing the model to varied expressions of identical knowledge encourages it to focus on the underlying logical structure of the information, rather than its superficial presentation. This could foster a more direct link between an individual’s name and their attributes. We will introduce probing techniques to substantiate this hypothesis in Section 5.

#### 4.3 CELEBRITY CAN HELP MINORITY

The previous subsection highlighted the significant benefits of knowledge augmentation. However, in practice, we may not have augmented data for all individuals. This subsection explores whether partially augmenting data can improve knowledge extraction for non-augmented data. In our biography dataset, the augmented subset is akin to a “celebrity” group with plentiful online biographical information, potentially included in the fine-tuning dataset as well. The non-augmented subset is comparable to a “minority” group with limited biographical data.

For comparison, we introduce an additional set of  $N = 100,000$  individuals, the celebrity group  $\mathcal{P}_{\text{cel}}$ , while the original  $N$  individuals form the minority group  $\mathcal{P}_{\text{min}}$ . We test both synthetic `bioS` and more realistic `bioR` data. For `bioS`, the celebrity group’s biographies use the `multi5+permute` augmentation, simulating varied expressions found on internet. For `bioR`, the celebrity group uses the `multi5` augmentation, generating their biographies five times using LLaMA.

The language model is pretrained on the combined set  $\mathcal{P}_{\text{cel}} \cup \mathcal{P}_{\text{min}}$  biographies and then fine-tuned using QAs from the celebrity group  $\mathcal{P}_{\text{cel}}$ . We evaluate the model’s QA accuracy on the  $\mathcal{P}_{\text{min}}$  group.<sup>10</sup> Our results are presented in Figure 4.

<sup>7</sup>For `bioS` data, each of the six sentences is selected from around 50 templates, with a new template resampled for each sentence in the  $M$  entries. For `bioR` data, we recreate the biography using LLaMA for each of the  $M$  entries.

<sup>8</sup>For `bioS single`, we denote random permutation of the same six sentences  $P$  times as `permuteP`. For `bioS multiM`, we denote random permutation of each of the  $M$  biography entries as `permute`. The `bioR` data, generated by LLaMA, already has some randomness in sentence ordering, so no extra permutations are added.

<sup>9</sup>An exception is when permutation is directly added to the single data without multiplicity (see “`bioS single + permute1`”), this hurts the QA performance as it makes knowledge extraction harder.

<sup>10</sup>Other fine-tuning variations, such as QA fine-tuning with half of  $\mathcal{P}_{\text{min}}$  as training and half as testing, show negligible differences.



baseline	2.7	0.0	0.5	0.3	1.0	0.4	13.7
bioS single + permute1	4.4	0.5	3.3	2.4	5.0	3.5	13.7
bioS single + permute1 + CEL	86.8	98.3	96.8	90.7	90.2	71.7	80.1
bioR single	10.0	25.1	13.9	2.4	5.5	2.0	14.1
bioR single + wiki	7.3	18.4	5.2	2.6	4.3	1.8	14.1
bioR single + CEL	76.3	94.3	85.3	82.9	79.4	67.0	56.6

Figure 4: QA finetune accuracy on the *minority group* with vs. without celebrity data in the pretraining process. Experiment details are in Appendix I, where we also include additional experiments in Figure 16.

**Results.** In the synthetic bioS case, introducing celebrity data boosts the minority group’s QA accuracy from 4.4% to 86.8%. This is significant because:

- the minority group’s BIO pretrain data *remains unchanged* in both cases, with  $\mathcal{P}_{\min}$  using bioS single+permute1 for biographies, and
- the minority group’s QA data *is not used* during fine-tuning.

This highlights that **simply including celebrity data during pretraining** significantly improves the model’s ability to store and extract knowledge from the minority group. Similarly, in the more realistic bioR case, introducing celebrity data also increases the minority group’s QA accuracy from 10.0% to 76.3%. We believe this strongly suggests that this phenomenon *also occurs in real-world scenarios*. We will introduce probing techniques to validate the above findings in Section 5.

*Remark 4.1.* Using the bioR dataset, we find the positive impact of celebrity data is *not universal*. Substituting it with the WikiBook dataset improves the model’s English comprehension, yet it still struggles with biographical knowledge extraction. This suggests that only celebrity data of *similar form* truly aids knowledge extraction for minority groups. In Figure 16 in Appendix I, we further investigate different celebrity data types and instances of minor format differences between minority and celebrity knowledge.

## 5 KNOWLEDGE PROBES ON THE BIO PRETRAINED MODEL

We investigate how a pretrained language model on BIO data encodes knowledge in its hidden representations using two probing techniques: position-based probing (P-probing) and query-based probing (Q-probing). Both techniques employ simple (nearly-linear) probes to extract a person’s attributes from the model’s hidden representations. Detailed findings are in Appendix A.

**In P-probing,** we input biography entries into the pretrained model and train a linear classifier on the last hidden layer to predict six target attributes. To accommodate varied data lengths, we identify six *special token positions* preceding the first occurrences of the six attributes in each biography entry. We use the transformer’s last hidden layer at these positions to (linearly) predict the six target attributes (Figure 5).<sup>11</sup> Our results (Figure 6) show that *increased knowledge augmentation* in the pretrain data improves *P-probing prediction accuracies from earlier token positions*. In the basic bioS single setup, P-probing accuracy remains low until the token immediately preceding the target attribute. This suggests the model memorizes BIO data but encodes knowledge in a complex manner, revealing a person’s attribute **only after encountering all prior attributes**. This **prevents knowledge extraction during QA finetuning**, particularly when only the person’s name is given. In Appendix A, we use a Venn diagram to precisely illustrate which attribute is stored after observing another, further confirming this finding.

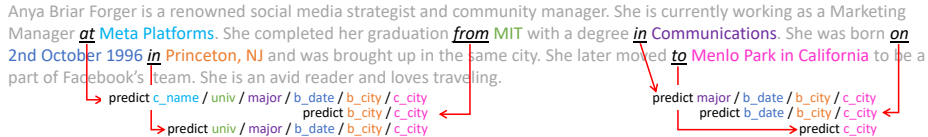


Figure 5: Illustration of the P-probing. Underscore prepositions are the *special token positions* where we prob. The task is to predict all attributes following these positions. Given the attribute ordering, there can be up to  $6 \times 6 = 36$  tasks across all data.

<sup>11</sup>For each target attribute prediction task, we freeze the pretrained network but add a trainable rank-2 update on the embedding layer to account for the task change.



baseline	8.3	8.3	8.3	8.3	8.3	2.5	2.5	2.5	2.5	2.5	2.5	37.0	37.0	37.0	37.0	37.0	4.0	4.0	4.0	4.0	4.0	4.0	1.5	1.5	1.5	1.5	1.5	1.5	14.8	14.8	14.8	14.8	14.8	14.8
bioS single	100	100	100	100	100	5.9	100	100	100	100	100	38.0	37.1	99.2	100	100	4.7	4.6	5.4	99.9	100	100	1.5	1.2	1.3	2.4	99.5	100	15.4	15.4	14.9	13.0	69.1	100
bioS single + fullname	100	100	100	100	100	5.9	100	100	100	100	100	38.0	37.1	99.2	100	100	4.7	4.6	5.4	99.9	100	100	1.5	1.2	1.3	2.4	99.5	100	15.4	15.4	14.9	13.0	69.1	100
bioS single + permute1	100	100	100	100	100	5.9	100	100	100	100	100	38.0	37.1	99.2	100	100	4.7	4.6	5.4	99.9	100	100	1.5	1.2	1.3	2.4	99.5	100	15.4	15.4	14.9	13.0	69.1	100
bioS single + permute2	100	100	100	100	100	5.9	100	100	100	100	100	38.0	37.1	99.2	100	100	4.7	4.6	5.4	99.9	100	100	1.5	1.2	1.3	2.4	99.5	100	15.4	15.4	14.9	13.0	69.1	100
bioS single + permute5	100	100	100	100	100	5.9	100	100	100	100	100	38.0	37.1	99.2	100	100	4.7	4.6	5.4	99.9	100	100	1.5	1.2	1.3	2.4	99.5	100	15.4	15.4	14.9	13.0	69.1	100
bioS single + permute1 + fullname	100	100	100	100	100	5.9	100	100	100	100	100	38.0	37.1	99.2	100	100	4.7	4.6	5.4	99.9	100	100	1.5	1.2	1.3	2.4	99.5	100	15.4	15.4	14.9	13.0	69.1	100
bioS single + permute2 + fullname	100	100	100	100	100	5.9	100	100	100	100	100	38.0	37.1	99.2	100	100	4.7	4.6	5.4	99.9	100	100	1.5	1.2	1.3	2.4	99.5	100	15.4	15.4	14.9	13.0	69.1	100
bioS single + permute5 + fullname	100	100	100	100	100	5.9	100	100	100	100	100	38.0	37.1	99.2	100	100	4.7	4.6	5.4	99.9	100	100	1.5	1.2	1.3	2.4	99.5	100	15.4	15.4	14.9	13.0	69.1	100
bioS multi2	100	100	100	100	100	5.9	100	100	100	100	100	38.0	37.1	99.2	100	100	4.7	4.6	5.4	99.9	100	100	1.5	1.2	1.3	2.4	99.5	100	15.4	15.4	14.9	13.0	69.1	100
bioS multi2 + fullname	100	100	100	100	100	5.9	100	100	100	100	100	38.0	37.1	99.2	100	100	4.7	4.6	5.4	99.9	100	100	1.5	1.2	1.3	2.4	99.5	100	15.4	15.4	14.9	13.0	69.1	100
bioS multi2 + permute1	100	100	100	100	100	5.9	100	100	100	100	100	38.0	37.1	99.2	100	100	4.7	4.6	5.4	99.9	100	100	1.5	1.2	1.3	2.4	99.5	100	15.4	15.4	14.9	13.0	69.1	100
bioS multi2 + permute2	100	100	100	100	100	5.9	100	100	100	100	100	38.0	37.1	99.2	100	100	4.7	4.6	5.4	99.9	100	100	1.5	1.2	1.3	2.4	99.5	100	15.4	15.4	14.9	13.0	69.1	100
bioS multi2 + permute5	100	100	100	100	100	5.9	100	100	100	100	100	38.0	37.1	99.2	100	100	4.7	4.6	5.4	99.9	100	100	1.5	1.2	1.3	2.4	99.5	100	15.4	15.4	14.9	13.0	69.1	100
bioS multi2 + permute1 + fullname	100	100	100	100	100	5.9	100	100	100	100	100	38.0	37.1	99.2	100	100	4.7	4.6	5.4	99.9	100	100	1.5	1.2	1.3	2.4	99.5	100	15.4	15.4	14.9	13.0	69.1	100
bioS multi2 + permute2 + fullname	100	100	100	100	100	5.9	100	100	100	100	100	38.0	37.1	99.2	100	100	4.7	4.6	5.4	99.9	100	100	1.5	1.2	1.3	2.4	99.5	100	15.4	15.4	14.9	13.0	69.1	100
bioS multi2 + permute5 + fullname	100	100	100	100	100	5.9	100	100	100	100	100	38.0	37.1	99.2	100	100	4.7	4.6	5.4	99.9	100	100	1.5	1.2	1.3	2.4	99.5	100	15.4	15.4	14.9	13.0	69.1	100
bioS multi5	100	100	100	100	100	5.9	100	100	100	100	100	38.0	37.1	99.2	100	100	4.7	4.6	5.4	99.9	100	100	1.5	1.2	1.3	2.4	99.5	100	15.4	15.4	14.9	13.0	69.1	100
bioS multi5 + fullname	100	100	100	100	100	5.9	100	100	100	100	100	38.0	37.1	99.2	100	100	4.7	4.6	5.4	99.9	100	100	1.5	1.2	1.3	2.4	99.5	100	15.4	15.4	14.9	13.0	69.1	100
bioS multi5 + permute1	100	100	100	100	100	5.9	100	100	100	100	100	38.0	37.1	99.2	100	100	4.7	4.6	5.4	99.9	100	100	1.5	1.2	1.3	2.4	99.5	100	15.4	15.4	14.9	13.0	69.1	100
bioS multi5 + permute2	100	100	100	100	100	5.9	100	100	100	100	100	38.0	37.1	99.2	100	100	4.7	4.6	5.4	99.9	100	100	1.5	1.2	1.3	2.4	99.5	100	15.4	15.4	14.9	13.0	69.1	100
bioS multi5 + permute5	100	100	100	100	100	5.9	100	100	100	100	100	38.0	37.1	99.2	100	100	4.7	4.6	5.4	99.9	100	100	1.5	1.2	1.3	2.4	99.5	100	15.4	15.4	14.9	13.0	69.1	100
bioS multi5 + permute1 + fullname	100	100	100	100	100	5.9	100	100	100	100	100	38.0	37.1	99.2	100	100	4.7	4.6	5.4	99.9	100	100	1.5	1.2	1.3	2.4	99.5	100	15.4	15.4	14.9	13.0	69.1	100
bioS multi5 + permute2 + fullname	100	100	100	100	100	5.9	100	100	100	100	100	38.0	37.1	99.2	100	100	4.7	4.6	5.4	99.9	100	100	1.5	1.2	1.3	2.4	99.5	100	15.4	15.4	14.9	13.0	69.1	100
bioS multi5 + permute5 + fullname	100	100	100	100	100	5.9	100	100	100	100	100	38.0	37.1	99.2	100	100	4.7	4.6	5.4	99.9	100	100	1.5	1.2	1.3	2.4	99.5	100	15.4	15.4	14.9	13.0	69.1	100

Figure 6: P-probing accuracies for various pretrained models on bioS data. Each **row** represents a pretrained model using a different knowledge augmentation, and each **column** labeled “*i*-field” shows the accuracy of predicting the *first token* of *field* from position *i*. Details are in Section 5 and Appendix G (where we also include experiments for the bioR data and for predicting the full-attribute *field*.)

baseline	2.7	0.0	0.5	0.3	1.0	0.4	13.7	8.3	2.5	37.0	4.0	1.5	14.8	0.5	0.3	1.0	0.4	13.7
bioS single	9.7	33.5	6.3	2.3	4.0	1.1	13.8	63.4	1.9	37.5	3.1	0.2	13.1	1.1	0.3	1.4	0.1	11.6
bioS single + fullname	48.9	56.2	58.8	63.0	55.7	50.5	14.1	78.8	47.5	65.4	51.0	47.4	28.9	43.9	31.2	40.8	44.4	28.9
bioS single + permute1	4.4	0.5	3.3	2.4	5.0	3.5	13.7	10.2	1.2	37.5	2.9	0.7	12.8	0.5	0.4	1.5	0.6	12.0
bioS single + permute2	52.8	57.3	48.3	53.1	55.0	51.8	14.8	85.4	42.1	60.4	58.7	53.8	63.1	79.6	30.9	48.2	93.3	58.8
bioS single + permute5	70.0	56.4	57.7	58.3	64.9	90.5	9.7	95.3	49.7	66.6	68.6	96.5	99.1	47.7	40.7	63.0	96.2	98.9
bioS single + permute1 + fullname	31.7	26.6	29.3	36.9	31.1	31.4	37.9	50.1	20.9	42.6	27.1	27.3	34.7	19.3	13.5	21.2	26.8	33.4
bioS single + permute2 + fullname	73.1	69.0	60.6	64.2	64.0	87.9	95.0	77.8	51.0	63.3	61.4	92.1	96.7	48.1	35.0	51.5	91.9	96.6
bioS single + permute5 + fullname	80.2	83.7	67.8	72.6	69.1	93.0	98.6	86.5	59.0	73.5	64.2	98.0	99.9	56.8	40.8	56.6	97.9	99.9
bioS multi2	41.0	100	71.7	95.1	76.1	52	14.0	100	70.3	45.1	19.6	0.7	13.0	68.8	14.9	14.8	0.7	11.8
bioS multi2 + fullname	86.0	100	97.7	99.5	97.6	91.5	13.5	94.5	82.2	84.2	92.7	99.8	99.2	94.5	82.2	84.2	92.7	99.8
bioS multi2 + permute1	91.2	99.3	98.7	99.8	96.7	83.3	13.5	99.9	99.9	99.6	99.4	93.9	90.2	98.2	78.6	95.5	93.1	89.4
bioS multi2 + permute2	96.1	100	98.8	91.3	98.1	93.7	97.8	99.7	100	100	99.9	99.9	99.4	95.4	84.6	88.1	99.4	99.1
bioS multi2 + permute5	41.0	100	90.8	30.9	43.5	10.2	13.8	100	89.8	40.3	10.1	3.2	13.2	38.1	9.7	36.1	2.3	11.9
bioS multi5	82.4	100	98.6	88.4	96.1	91.9	26.8	100	100	98.3	99.4	98.5	13.1	95.2	58.7	87.1	194.1	66.9
bioS multi5 + fullname	96.6	100	99.0	91.3	97.7	95.1	19.7	100	100	99.9	100	99.9	99.8	96.1	72.6	94.9	99.6	99.7
bioS multi5 + permute1	96.2	100	98.7	90.6	97.9	93.7	99.0	100	100	99.8	100	100	99.8	95.4	88.1	92.0	99.9	99.8
bioS multi5 + permute2																		
bioS multi5 + permute5																		

Figure 7: Q-probing accuracies. Each **row** denotes a pretrained model with its specific knowledge augmentation. The left block reiterates QA finetune accuracies from Figure 3. The middle showcases Q-probing accuracies on the first-token prediction for the six attributes, and the right focuses on Q-probing for the whole-attribute prediction. (Further details for bioR and more are in Appendix G. Note: For birth date, first token predicts the whole birth month; we do not have whole-attribute prediction for it since it has too many choices.)

**In Q-probing**, we focus on the knowledge directly linked to a person’s name. We evaluate input sentences *containing only* the person’s full name and train a linear classifier on the last layer’s hidden states to predict the person’s six attributes.<sup>12</sup> Our results (Figure 7 in Appendix A.2) show that the knowledge-extraction finetune accuracy is *directly linked* to whether the knowledge is (nearly-)linearly stored on the person’s name in the pretrained model. This is a property independent of the finetune parameters, and suggests that the model *does not utilize contextual or global information from the biographies to extract knowledge about the individual*.

## 6 CONCLUSION

This study explores the ability of pre-trained language models to store and extract knowledge during inference using question-answering tasks. We created a semi-synthetic biography dataset and utilized probing techniques to examine the effect of knowledge augmentation on the storage and extractability of knowledge in pre-trained transformers. Synthetic data offers increased control over model training and fine-tuning inputs, which is crucial for understanding the influence of different data sources on the **internal mechanisms** of transformers. This could potentially be a significant future direction for unraveling the complexities of transformers. The paper also highlights the **importance of rewriting** essential but infrequently occurring data during pre-training to ensure its effective storage for subsequent tasks. This should be achieved using tools like ChatGPT before pre-training, as rectification during the fine-tuning stage might be too late if the pre-training data has not been fully augmented. While our primary focus was on autoregressive language models, our techniques are also applicable to bidirectional models like BERT, as discussed in Appendix J.

<sup>12</sup>We freeze all transformer layers (acquired through pretraining), except the embedding layer, to which we apply a rank-16 update. This adjustment is arguably the minimal change necessary since we are tackling a notably different input distribution.

## REFERENCES

- Anonymous. Knowledge Storage and Extraction in Language Models (Part B). *submitted to ICLR 2024*, September 2023. see also supplementary materials of this paper.
- Carlos Aspillaga, Marcelo Mendoza, and Alvaro Soto. Inspecting the concept knowledge graph encoded by modern language models. *arXiv preprint arXiv:2105.13471*, 2021.
- Lukas Berglund, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomasz Korbak, Daniel Kokotajlo, and Owain Evans. Taken out of context: On measuring situational awareness in llms. *arXiv preprint arXiv:2309.00667*, 2023.
- Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of the ACL Workshop on Challenges & Perspectives in Creating Large Language Models*, 2022. URL <https://arxiv.org/abs/2204.06745>.
- Hengyi Cai, Hongshen Chen, Yonghao Song, Cheng Zhang, Xiaofang Zhao, and Dawei Yin. Data manipulation: Towards effective instance learning for neural dialogue generation via learning to augment and reweight. *arXiv preprint arXiv:2004.02594*, 2020.
- Byeongmin Choi, YongHyun Lee, Yeunwoong Kyung, and Eunchan Kim. Albert with knowledge graph encoder utilizing semantic similarity for commonsense question answering. *arXiv preprint arXiv:2211.07065*, 2022.
- Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*, 2018.
- Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*, 2021.
- Ronen Eldan and Yuanzhi Li. Tinstories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*, 2023.
- Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*, 2020.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.
- Evan Hernandez, Belinda Z Li, and Jacob Andreas. Measuring and manipulating knowledge representations in language models. *arXiv preprint arXiv:2304.00740*, 2023.
- Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2021.
- Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.
- Sosuke Kobayashi. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 452–457, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2072. URL <https://aclanthology.org/N18-2072>.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020. URL [https://proceedings.neurips.cc/paper\\_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf).
- Belinda Z Li, Maxwell Nye, and Jacob Andreas. Implicit representations of meaning in neural language models. *arXiv preprint arXiv:2106.00737*, 2021.

- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv e-prints*, abs/1907.11692, July 2019.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.
- Tahira Naseem, Srinivas Ravishankar, Nandana Mihindukulasooriya, Ibrahim Abdelaziz, Young-Suk Lee, Pavan Kapanipathi, Salim Roukos, Alfio Gliozzo, and Alexander Gray. A semantics-aware transformer model of relation linking for knowledge base question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 256–262, Online, August 2021. Association for Computational Linguistics.
- Reham Omar, Omij Mangukiya, Panos Kalnis, and Essam Mansour. Chatgpt versus traditional question answering for knowledge graphs: Current status and future directions towards knowledge graph chatbots. *arXiv preprint arXiv:2302.06466*, 2023.
- Hao Peng, Xiaozhi Wang, Shengding Hu, Hailong Jin, Lei Hou, Juanzi Li, Zhiyuan Liu, and Qun Liu. Copen: Probing conceptual knowledge in pre-trained language models. *arXiv preprint arXiv:2211.04079*, 2022.
- Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.
- Kyle Richardson and Ashish Sabharwal. What does my QA model know? devising controlled probes using expert knowledge. *Transactions of the Association for Computational Linguistics*, 8:572–588, 2020. doi: 10.1162/tacl.a.00331. URL <https://aclanthology.org/2020.tacl-1.37>.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*, 2022.
- Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2021.
- Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. Head-to-tail: How knowledgeable are large language models (llm)? aka will llms replace knowledge graphs? *arXiv preprint arXiv:2308.10168*, 2023.
- Madhumita Sushil, Simon Suster, and Walter Daelemans. Are we there yet? exploring clinical domain knowledge of BERT models. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pp. 41–53, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.bionlp-1.5. URL <https://aclanthology.org/2021.bionlp-1.5>.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*, 2023.
- Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

# APPENDIX

## A KNOWLEDGE PROBES ON THE BIO PRETRAINED MODEL

We investigate how a language model, *pretrained on BIO data*, encodes knowledge in its hidden states. We propose two probing methods: position-based (P-probing) and query-based (Q-probing). Both methods employ simple, nearly-linear probes to extract personal attributes from the model’s hidden states.

### A.1 POSITION-BASED PROBING

In P-probing, we feed biography entries into a pretrained model and train an additional linear classifier on the model’s final hidden layer to predict six target attributes (e.g., university, major, etc.). This probing provides insight into how these attributes are encoded during pretraining.

The model, already well-pretrained on biography data, is expected to encode the target attribute with 100% accuracy at positions immediately preceding the attribute. However, our goal is to determine if the model retains this information at positions far before the corresponding attribute.

For instance, if the linear classifier shows high accuracy right after the person’s full name, it implies that the model is directly learning “Anya’s employer is Meta Platforms”. If high accuracy is only achieved at the biography’s end, the model might be using flawed logic, such as “the birthday is October 2, 1996, the university is MIT, hence the employer is Meta.”

**Our P-probing technique.** To manage data of varying lengths, we identify six *special token positions* immediately before the first occurrences of the six attributes in each biography entry. We then train a linear classifier on the transformer’s final hidden layer at these positions to predict each of the six target attributes.<sup>13</sup> Refer to Figure 5 for an illustration. This results in  $6 \times 6$  classification tasks. For each prediction task, we freeze the entire pretrained network but add a trainable rank-2 update on the embedding layer to accommodate the task change.<sup>14</sup>

#### A.1.1 P-PROBING MAIN RESULTS

Our results, shown in Figure 6 and further detailed for the bioR data in Appendix G, indicate that P-probing prediction accuracies (on the BIO-pretrained model) are *typically higher* at earlier special positions in models with *high out-of-distribution QA accuracy* (after QA finetuning). Specifically,

- In the bioS single setup, P-probing accuracy remains low until the token immediately preceding the target attribute. The accuracy is around 2% when predicting the company name from earlier token positions, but it increases to 99.5% when evaluating on the token position right before the company name. This suggests that the model memorizes all the BIO data during pretraining, but encodes knowledge in a complex manner, only revealing a person’s attribute *after encountering all prior attributes for that individual*. This **prevents knowledge extraction** during QA finetuning, especially when only the person’s name is provided.
- In the heavily augmented setup like bioS multi5+permute, the P-probing accuracy for all six attributes rises to nearly 100% from the first special position, which is before *all* of the attributes. This indicates that the model not only memorizes the BIO data but also identifies the person’s complete attributes solely upon seeing the person’s name, **facilitating knowledge extraction** during the QA finetuning process.
- For intermediate setups, the results are mixed. For example, comparing bioS single with multi5, we find that adding multiplicity (without permutation) results in earlier attribute storage, accounting for the increase in QA finetune accuracy from 9.7% to 41% as seen in Figure 3. Similarly, comparing bioS single+permute1 with single+permute5, we observe that permuting the six sentences five times (without diversifying the sentences) also leads to earlier knowledge storage, explaining the rise in QA finetune accuracy from 4.4% to 70%.

<sup>13</sup>For GPT2-small with 768 hidden dimensions, this trainable linear classifier is of dimension  $768 \times M$  for each target attribute with  $M$  possibilities, regardless of the context length.

<sup>14</sup>For GPT2-small with 768 hidden dimensions and vocab size  $V$ , this rank-2 update has  $2V + 2 \times 768$  trainable parameters. More details can be found in Appendix G.

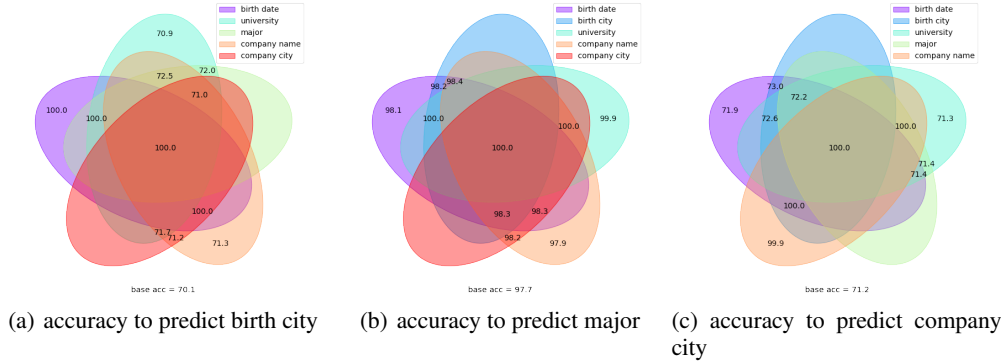


Figure 8: Closer P-probing on bioS couple data. The Venn diagram shows prediction accuracy for the target attribute at specific token positions, based on whether each of the remaining five attributes has been seen or not. **Observation:** Predicting “birth city” improves with “birth date”; “major” with “university”; and “company city” with “company name.”

#### A.1.2 CLOSER P-PROBING AT KNOWLEDGE DEPENDENCY

As noted earlier, the model may infer attribute relationships based on their order in pretraining data. For instance, if a birth date always precedes a company city, the model might infer “the person born on October 2, 1996 works in Menlo Park” instead of “Anya’s work city is Menlo Park”. This can occur if the pretraining data isn’t adequately augmented, and the model may even favor linking one attribute to another, rather than to the person’s name, if two attributes are closely correlated (such as company city and company name).

To investigate this, we created a variant of the bioS dataset, grouping the 6 sentences into 3 pairs with a consistent order: birthdate before birth city, university before major, and work company before work city. We allowed random permutations among these pairs and sentence diversities. We refer to this dataset as bioS couple (see Appendix C for details).

We examined our P-probing on this dataset as  $2^5 \times 6$  classification tasks, predicting each of the six target attributes from a special token position where only a subset  $S$  of the remaining five attributes has been observed ( $S$  has  $2^5$  possibilities).<sup>15</sup> Our results, visualized in Figure 8, show that the accuracy in predicting the second attribute in each pair is heavily influenced by whether the model has encountered the first attribute, even with substantial data diversity.

*Remark A.1.* This observation relates to Figure 3, where the “company city” attribute consistently shows the weakest QA finetune performance in the bioS dataset family. This is due to our data construction approach, where “company city” is determined solely by “company name”. The model thus associates “company city” with “company name” rather than the person’s name, especially if the company name is presented earlier. This explains why, in Figure 3, the model shows virtually “zero” prediction accuracy for an individual’s company city using only their name.<sup>16</sup>

#### A.1.3 P-PROBING EXTENSIONS

We could consider alternative P-probing forms, such as introducing a low-rank update to the pre-trained model’s main body, like a trainable LoRA update with a small rank on the query/value matrices. While not necessary for our positive results (e.g., the highly augmented data bioS multi5+permute), it could be interesting to apply this to the negative results (e.g., the basic data bioS single). However, our experiments showed no significant increase in P-probing accuracies, so we omit the details.

Our P-probing has focused on the six distinct token positions, likely the preposition words preceding the six attributes. How about probing other positions, like tokens following each attribute or the

<sup>15</sup>The P-probing process remains the same as before, using only 6 sets of trainable parameters each for a target attribute, each with a single classification linear layer and a single rank-2 update on the embedding. The difference is a more detailed interpretation of the results.

<sup>16</sup>This dynamic can be explored as a form of *knowledge manipulation*. For example, if the language model is good at retaining work company names, can it determine work city locations as a simple classification task using just company names? We explore this in our parallel paper Anonymous (2023).

person’s name? We observed that P-probing accuracy might improve as the model processes more “extraneous” tokens. For instance, the P-probing accuracy for a person’s birth date could increase after encountering phrases like “was born on” or “has birthday in”. This could be due to the model’s ability to associate the birthdate information with the sentence’s *structure*. We chose not to include these observations for clarity.

In Appendix G, we demonstrate the difference between a rank-2 and a rank-4 update on the embedding layer. The results confirm that a rank-2 update is sufficient for P-probing on our biography data.

## A.2 QUERY-BASED PROBING

P-probing offers a qualitative assessment of early knowledge storage in the model relative to the original biography entry. However, it can be limiting due to its dependence on the exact context structure from the biography entry. For instance, in Figure 5, knowledge may be stored in short phrases like “received mentorship and guidance.”

In query-based probing (Q-probing), we aim for a more precise, context-free value from a pretrained model, focusing on the knowledge directly associated with a person’s name. We evaluate sentences containing only the person’s full name and train a linear classifier on the last layer’s hidden states to predict the person’s six attributes. High accuracy suggests that the model directly links each person’s attributes to their name.

We consider an input sentence containing only the person’s full name, preceded by a starting token and followed by an ending token. Like P-probing, we freeze all transformer layers (acquired through pretraining), except the embedding layer, where we apply a low-rank update (using rank 16, compared to rank 2 in P-probing). This minimal change is necessary as we are addressing a distinct classification task under a different input distribution. We extract the hidden states from the last layer on the ending token and place a trainable linear classifier on top to predict the person’s six attributes. More details are in Appendix H.

**Our findings.** Our results are in Figure 7. Our main findings are:

- The QA finetune accuracy correlates closely with Q-probing accuracy, indicating that the *degree to which the attribute is directly linked to the person’s name* is a **crucial factor for effective knowledge extraction**. If the model fails to store knowledge properly during pretraining, QA finetuning may not rectify this.
- After applying knowledge augmentations to the pretraining data, Q-probing accuracy significantly increases. This suggests that the model encodes knowledge almost linearly in the hidden states directly adjacent to the person’s name. Thus, the linear probes can extract the person’s attributes from these hidden states as effectively as the model can be adapted through QA finetuning to answer questions related to those attributes.

In conclusion, Q-probing shows that language models create a direct mapping from a person’s name to their attributes. At the last hidden-layer, the model neither uses complex or nonlinear transformations nor leverages interactions between hidden states at different token positions to store and extract knowledge about the person. This implies that the model **does not use contextual or global information from the biographies to extract knowledge about the individual**.

## A.3 PROBING ON THE CELEBRITY AUGMENTATION

We also use P-probing and Q-probing to assess the impact of including celebrity data on knowledge storage and extraction for minority groups. This addition enhances the model’s storage efficiency, *even for* minority groups, implying that their attributes are more directly and closely associated with their names. These results are detailed in Figure 17 and Figure 18 in Appendix I.

## B KNOWLEDGE STORAGE FOR BIDIRECTIONAL MODELS

This paper primarily explores knowledge storage and extraction in language modeling through an autoregressive task. One may argue that some knowledge issues, such as the consistent knowledge ordering in bioS single, are unique to this task due to its *unidirectional* nature. We thus pose the question, *Could BERT be a solution to this?*

We analyze the BERT model Kenton & Toutanova (2019), similar to GPT2 but with a full attention matrix, allowing every token to attend to every other token. For a direct comparison, we modify our GPT2 architecture to replace its triangular attention matrix with a full matrix, keeping the GPT2 tokenizer and rotary embedding. We call this modified model GBERT.

Our pretraining task is now *whole-word masked-language modeling (MLM)*. Each English whole-word has a 15% chance of being selected, which is then replaced with a <MASK> token (80% chance), retained (10% chance), or replaced with a random token (10%). The goal is to predict the original word for these selected tokens.

For GBERT, we modify the QA task to evaluate its knowledge extraction capabilities. For questions like “What is the birth city of Anya Briar Forger?”, we append them with several <MASK> tokens (equaling the answer’s length).<sup>17</sup> A correct answer requires accurate recovery of all masked tokens.

We display results for both mixed training and BIO pretrain + QA finetune. Half of the QAs are used for mixed training (or QA fine-tuning), while we test out-of-distribution generalization accuracies on QAs for the remaining half of the people. Q-probing results for GBERT are also presented, determining if the model, with minor embedding layer modifications, can linearly predict target attributes from a person’s full name.

**Our findings.** Our findings are displayed in Figure 9. Key observations include:

- The QA-finetune and Q-probing accuracies show a strong correlation. This suggests that the ability to extract knowledge from a BERT-like model depends on whether such information is nearly linearly stored in hidden states directly adjacent to the person’s name. This reinforces the idea that our Q-probing technique might be effective for both decoder models like GPT2 and encoder models like BERT.
- Consistent with Figure 3, mixed training yields slightly superior out-of-distribution QA accuracies compared to BIO pretrain + QA finetune.
- Interestingly, the model performs well on “birth date” and “major” attributes but struggles on others. The reason is simple. In MLM, where each word has an equal chance of being masked, the model learns to associate knowledge words with the *most related* unmasked word, preferably those that are *adjacent*. For instance, words representing the “birth date” attribute (month, day, year) are quite independent, making the model more inclined to link them to the person’s name. For attributes like birth city, where there’s a strong link between the city “Bellevue” and state “WA”, the model maximizes this association, *inhibiting storage of knowledge on person names*.<sup>18</sup>

In conclusion, while bidirectional models like BERT might seem less sensitive to the ordering of knowledge during MLM pretraining, **the MLM task doesn’t necessarily promote proper knowledge storage for subsequent extraction. Unless the knowledge is a standalone word or consists of independent words** (like month, day, year), extracting knowledge after MLM pretraining might prove challenging, if not impossible.

## C DETAILS ON DATA PREPARATION

### C.1 BIO DATASET BIOS

In the synthetic dataset labeled as bioS, we generate profiles for  $N = 100,000$  individuals. Each individual’s first, middle, and last names, birth date, birth city, university attended, major of study, and current employer are selected *independently* and randomly from a uniform distribution.

- First, middle, and last names are drawn from pools of 400, 400, and 1000 English names respectively. We apply rejection sampling to ensure all  $N$  individuals have unique full names.
- Birth years range from 1900 to 2099, months are selected from the 12 months, and days are chosen between 1 and 28.
- Birth cities are selected from 200 US cities, with their respective state abbreviations, such as Princeton, NJ and Cambridge, MA.

<sup>17</sup>Revealing the answer’s token count might seem unfair. However, given our aim to highlight GBERT’s limitations, this extra information doesn’t hinder our intentions.

<sup>18</sup>Similarly, many majors are single words so this explains its high QA test accuracy. In contrast, the words representing universities or company names/cities are more dependent.



Figure 9: Additional results on the GBERT model pretrained via masked language modeling (MLM). Mix training (left) versus BIO pretrain + QA finetune (middle left) versus Q-probing (middle right and right).

**Observation.** MLM doesn’t necessarily promote proper knowledge storage for subsequent extraction; unless the knowledge is a single word or consists of independent words (like month, day, year), extracting knowledge after MLM pretraining might still be nearly impossible. (Further details are in Appendix J, and note we have pretrained trained the model twice longer comparing to GPT.)

- Universities are drawn from a list of 300 US institutions. Some may have similar prefixes, like University of California, Berkeley/Irvine/Davis/etc.
- Majors are selected from 100 common college disciplines, including Computer Science, Physics, and Music.
- Employers are chosen from a list of 263 companies, featuring names like Meta Platforms, Microsoft, and Google.

Additionally,

- We introduce a “company city” attribute that *depends* on the US location of the employer’s headquarters. For instance, an employee of Meta would list Menlo Park, CA as their company city. Notably, 13.7% of the companies are headquartered in New York, NY. Thus, defaulting to New York, NY when predicting a person’s work city yields a base accuracy of 13.7%.

In the bioS dataset, we craft a biographical text entry for each individual, distilling their profile into six sentences. Each sentence illuminates a distinct attribute of the individual. To increase diversity, we select each sentence randomly from a set of pre-defined templates. Specifically, we have 46 sentence templates for birth dates, 49 for birth cities, 49 for universities, 52 for majors of study, 47 for employers, and 48 for company cities. Beyond (2.1), we provide several more examples below:

Carlos Jameson Stokes has his annual celebration on November 12, 2088. He celebrates his birth in San Francisco, CA. He graduated from Oklahoma State University. He explored the theoretical aspects of Information Systems. He contributed his expertise to United Airlines Holdings. He acquired industry knowledge while working in Chicago, IL.

Alondra Bennett Rooney celebrates their life journey every year on April 1, 1909. They owe their roots to Durham, NC. They benefited from the resources and facilities provided by University of South Alabama. They developed a strong foundation in Data Science. They had a job at The Southern Company. They were involved in the industry of Atlanta, GA.

Aidan Alexa Dennis’s birth is celebrated annually on July 17, 1968. She calls Palmdale, CA her birthplace. She specialized in her field of study at Stevens Institute of Technology. She completed a rigorous program in International Business. She had employment prospects at Johnson & Johnson. She gained work experience in New Brunswick, NJ.

(We assign a random pronoun (he/she/they) to each person.)<sup>19</sup>

In the basic configuration, we produce a *single biographical entry* for each individual, maintaining a consistent order for the six sentences as previously outlined. In average, a biographical entry has 73.0 tokens using GPT2 tokenization. We denote this configuration as “bioS single.” For comparison, we delve into 15 knowledge augmentations:

- bioS single+fullname: Pronouns are replaced with the person’s full name.
- bioS single+permute1/2/5: The six sentences in the biography entry are randomly permuted 1/2/5 times for each person. However, the full name only appears in the first sentence, with subsequent sentences using pronouns. This results in 1/2/5 biography entries for each person.
- bioS single+permute1/2/5+fullname: As with the previous augmentation, but the full name is used in all six sentences.

<sup>19</sup>Given that we are not employing a pretrained model sourced from the internet, we did not do fact-checking. For instance, a person’s major may not align with the business of the company they work for, and their birth year might largely precede the company’s establishment date.

- **bioS multi2/5**: 2 or 5 biographical entries are generated for each person, with each generation employing a re-sampled set of sentence templates.
- **bioS multi2/5+permute**: Building on **bioS multi2/5**, the six sentences within each biographical entry are randomly permuted. However, the full name appears only once in the first sentence.
- **bioS multi2/5+fullname**: Building on **bioS multi2/5**, pronouns are replaced with the individual’s full name across all sentences.
- **bioS multi2/5+permute+fullname**: Incorporating features from both **bioS multi2/5+permute** and **bioS multi2/5+fullname**, the pronouns are replaced with the individual’s full name and the six sentences are randomly permuted.

### C.1.1 BIOS COUPLE

In Section A.1.2, when delving deeper into P-probing, we also introduced a partial knowledge augmentation on the **bioS** dataset, which we termed **bioS couple**.

Specifically, we initially generate six sentences, each derived from a set of sentence templates similar to those in **bioS single**. We then group these six sentences into three pairs. The sentence describing a person’s birthdate always precedes the one discussing the person’s birth city. Similarly, the sentence detailing the person’s university consistently comes before the one about their major, and the one about their employer invariably precedes the sentence regarding their work city. Subsequently, we permute the order of these three pairs of sentences, resulting in  $3! = 6$  potential arrangements. The individual’s full name is restricted to appear only in the first sentence. For each individual, we create such a biographical entry five times, designating this dataset as **bioS couple**.

## C.2 BIO DATASET BIOR

We also examine a “close-to-real” dataset produced by LLaMA Touvron et al. (2023); Zhou et al. (2023). Specifically, for the previously mentioned set of  $N = 100,000$  individuals, we provide an instructive prompt to LLaMA, directing it to generate a biographical entry for us. To promote diversity in our results, we draw our instructive prompts from a collection of sentence templates. Three examples are presented below.

LLaMA prompt: Write a 50 words biography about a person whose name is Nicole Kevin Pratt. The person studied at Haverford College. The person studied Management there. The person was born and raised in Baltimore, MD. The person worked in Cincinnati, OH for Procter & Gamble. The person’s birthday falls on January 25, 1977.

LLaMA prompt: Write a 50 words biography about a person whose name is Hunter Bennett Kenny. The person attended Queens College, City University of New York for education. The person pursued a degree in Political Science there. The person originated from Augusta, GA. The person worked in Menomonee Falls, WI for Kohl’s. The person was born on March 25, 2033.

LLaMA prompt: Write a 50 words biography about a person whose name is Johnathan Charles Wade. The person attended Colorado State University for education. The person pursued a degree in Sociology there. The person was brought up in New York City, NY. The person worked in Northbrook, IL for Allstate. The person’s birthday falls on January 7, 2098.

Here are the generated outputs from LLaMA Touvron et al. (2023); Zhou et al. (2023):

Nicole Kevin Pratt is an American business executive. She is currently the Vice President of P&G Global Business Services at Procter & Gamble. She was born on January 25, 1977, in Baltimore, Maryland. She graduated from Haverford College with a degree in Management. P&G recruited her as an Assistant Brand Manager in 2000. She held various leadership positions in brand management, marketing, and sales across different business units and categories. She was named Vice President of P&G Global Business Services in 2019. Nicole currently lives in Cincinnati, Ohio with her husband and three children.

Hunter Bennett Kenny is a talented political science graduate from Queens College, City University of New York. He hails from Augusta, Georgia and was born on March 25, 2033. During his time at college, he was an active member of the student council and served as its president in his senior year. He interned at the office of New York Senator Chuck Schumer. After graduating cum laude, he worked for Kohl’s in Menomonee Falls, Wisconsin. He currently resides in Brooklyn, New York.

Johnathan Charles Wade is a successful insurance agent who works for Allstate. He was born on January 7, 2098, in New York City, NY. He graduated from Colorado State University, where he majored in Sociology. He currently resides in Northbrook, IL.

We observed that LLaMA might not always generate a biographical entry that includes all six attributes. To address this, we repeat the sampling process until LLaMA’s output ensures the inclusion of all attributes. Typically, the entry begins with the individual’s full name, and the order in which the six attributes appear can vary. In average, a biographical entry has 72.3 tokens using GPT2 tokenization.

In the basic configuration, we produce a single biographical entry for each person, denoted as “bioR single.” For comparison, we also introduce the multi $M$  augmentation, which creates  $M$  entries per person, and the fullname augmentation.

## D DETAILS ON MODEL ARCHITECTURE

The classic GPT2-small architecture Radford et al. (2019) comprises 12 layers, with 12 attention heads per layer, and has  $768 = 12 \times 64$  hidden dimensions (124M). Recent research Black et al. (2022); He et al. (2020); Su et al. (2021) has shown that transformers can achieve a significant performance improvement by utilizing attentions based on the *relative* positional differences of tokens. Consequently, in this paper, we omit the positional embedding from the GPT2 architecture and incorporate a rotary embedding, following the standard GPT-NeoX implementation Black et al. (2022) available on Huggingface (with the default frequency base of 10,000 and rotary dimension set to a quarter of the embedding dimension). We continue to refer to this as GPT2 for simplicity.

In our bioS experiments, we employ the above architecture. For the bioR experiments, we opt for a larger GPT model with 12 layers, 20 attention heads each 64-dimensional (302M), tailored to its increased difficulty.

Only when presenting our negative result in Figure 2, we also tried a 12-layer, 32-head (each 64-dimensional) GPT model (682M).

Additionally, we evaluate the BERT model Kenton & Toutanova (2019). BERT is similar to GPT2 but features a complete attention matrix, enabling every token to attend to all others. For a side-by-side comparison, we modify our GPT2 architecture to swap its triangular attention matrix for a full matrix, while keeping the GPT2 tokenizer and rotary embedding (removing positional embedding). We label this revised model GBERT. A primary distinction is that GBERT adopts pre-layernorm (inherited from the base GPT2 architecture), whereas BERT utilizes post-layernorm.

Throughout pretraining, mixed training, and QA finetuning, we maintain a context window length of 512.

## E DETAILS ON PRETRAIN AND MIX TRAINING

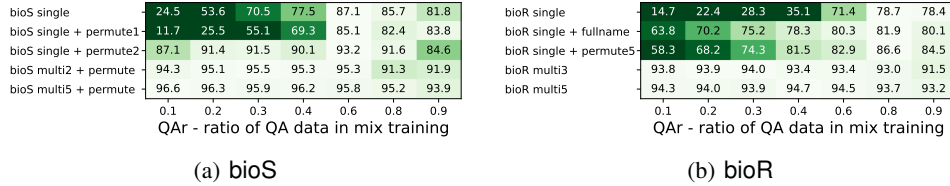


Figure 10: QA test accuracy for mix training across various choices of  $QA_r$ .

**Observation:** Particularly with more challenging data (i.e., with less knowledge augmentation), a greater  $QA_r$  ratio during training results in enhanced out-of-distribution QA (test) accuracy. This lends further credence to our assertion about the model’s unusual behavior: it initially acquires knowledge from QA rather than BIO and subsequently seeks to link BIO with QA.

During BIO pretraining, we randomly sample biographical entries of individuals and concatenate them to form sequences of 512 tokens, using a standard  $\langle EOS \rangle$  token to separate individual entries.

In mix training, we pre-train the model with BIO data from *all* individuals and QA data from *half* of them. Specifically, each training sequence of 512 tokens is either sourced entirely from the BIO entries (as previously mentioned) or entirely from the QA entries (again, from randomly sampled individuals and concatenated). We define a parameter  $QA_r$  to dictate the frequency of using QA entries. Predominantly in this paper, we set  $QA_r = 0.8$ , which implies a 2 : 8 ratio between BIO and QA entries in terms of the number of pre-trained tokens. We subsequently assess the model’s generation accuracy using QA data from the other half of the individuals. Refer to Figure 10 for an analysis of how the parameter  $QA_r$  impacts mix-training performance.

For both BIO pretraining and mix training, we employed a conventional set of optimization parameters: the AdamW optimizer with a weight decay of 0.1,  $\varepsilon = 10^{-6}$ , an initial learning rate of 0.001, a 1000-step linear warmup, and cosine learning rate decay (from 0.001 decreasing to 0.0001). We used a batch size of 96.

There were a total of 80,000 training steps for bioS (utilizing the GPT2 small architecture) and 150,000 training steps for bioR (which employs a larger 12-layer, 20-head GPT2). Only when using the 12-layer, 32-head GPT2 to present our negative result in Figure 2, we used 200,000 training steps.

*Remark E.1.* Our training time is long enough to ensure next-token prediction accuracy well above 99% for both BIO pretraining and mix training, when focusing on tokens describing six attributes per individual. These numbers are not included in this paper’s figures.

## F DETAILS ON QA FINETUNE

In our QA finetuning tasks, we first use a BIO pretrained model checkpoint and then apply either full finetuning or LoRA finetuning.

For full finetuning, we employ the AdamW optimizer with  $\varepsilon = 10^{-6}$ . We use weight decays of 0.01 and 0.001, and initial learning rates of 0.001, 0.0003, and 0.0001. There is no warmup, and we implement cosine learning rate scheduling (reducing to 10% of the initial learning rate), a batch size of 48, and a total of 50,000 training steps. Given that we are presenting a negative result for full finetuning (as seen in Figure 2), we display the best QA test accuracy among all the lr/wd parameter combinations.

For LoRA finetuning, we maintain the aforementioned AdamW configuration but set a consistent weight decay of 0.01 and an initial learning rate of 0.0003 for all tasks.

The results in Figure 11 suggest that for the purpose of QA finetuning, LoRA is generally a better option compared to full finetuning. While a large rank- $r$  update on the query/value matrices isn’t essential, it appears beneficial to have a significant rank- $r'$  update on the embedding layer to address the distribution shift from the BIO data to the QA data.

For this reason, in all subsequent experiments in this paper (notably Figure 3 and 12), when conducting QA finetuning, we use  $r' = 128$  and either  $r = 8$  or  $r = 16$ , presenting the best accuracy from the two runs.

## G DETAILS ON P-PROBING

In our P-probing experiments, we freeze the BIO pretrained GPT model and append a limited set of trainable parameters. Using the GPT2-small as an example, we introduce:

- a trainable rank-2 update for the embedding layer, having dimensions of  $50256 \times 2$  and  $2 \times 768$ ,
- for each prediction task that is an  $M$ -class classification problem, a trainable linear layer with dimensions of  $768 \times M$ ,
- preceding the linear layer, a layer normalization layer furnished with trainable affine parameters.

In the context of Q-probing, recall that we considered six classification sub-tasks (from 6 special locations) for every attribute prediction task. Specifically, for the birthdate attribute, we solely address its first-token prediction task, which is equivalent to predicting the individual’s birth month.<sup>20</sup> For the remaining five attributes, both the first-token and whole-attribute prediction tasks are examined. In sum, this results in 11 prediction tasks, each comprising 6 sub-tasks. For every one of these 11 tasks, we incorporate a distinct set of trainable parameters.

For optimization, the AdamW optimizer is employed with  $\varepsilon = 10^{-6}$ , weight decay of 0.3, an initial learning rate of 0.001, no warmup, and a linear learning rate decay (down to 0 in the end). We set the batch size of 50 and trained for 30,000 steps. During this P-probing training phase, we have turned on the dropout on the (frozen) pretrained GPT model to prevent overfitting.

<sup>20</sup>This is because a birthdate encompasses  $200 \times 12 \times 28$  potential choices, surpassing  $N/2$ , the number of training individuals.

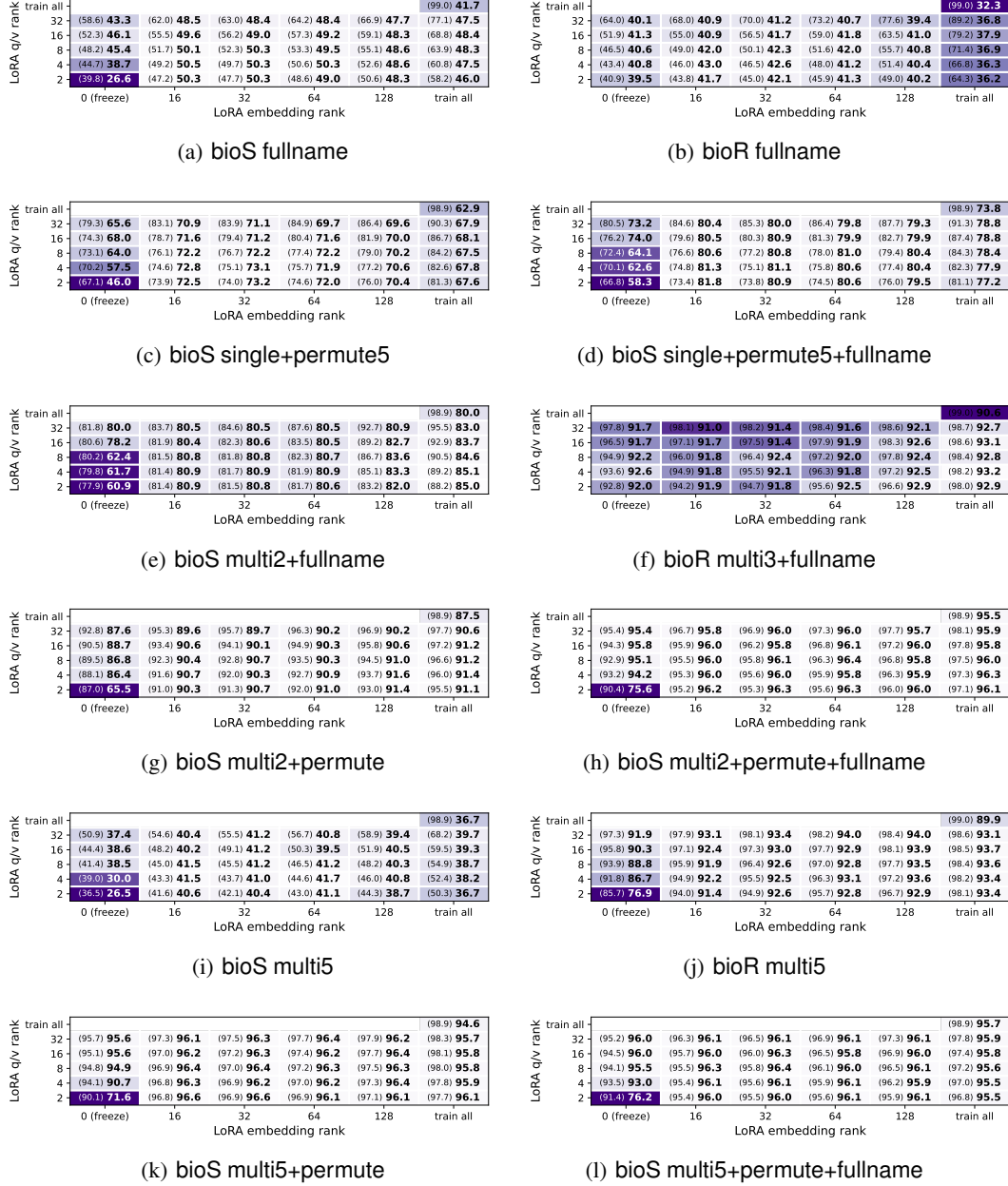


Figure 11: BIO pretrain + QA finetune (train acc) / test acc for various choices of fine-tuning settings. Bold number indicates QA generation accuracy on  $\mathcal{P}_{\text{test}}$ , and the smaller number in bracket represents QA (first-token) accuracy on  $\mathcal{P}_{\text{train}}$ . For LoRA fine-tune we consider a rank  $r = 2, 4, 8, 16, 32$  update on the query/value (q/v) matrices and a rank  $r' = 0, 16, 32, 64, 128$  update on the word embedding matrix. This is an extension of Figure 2.

baseline	2.7	0.0	0.5	0.3	1.0	0.4	13.7	2.7	0.0	0.5	0.3	1.0	0.4	13.7
bioR single	10.0	25.1	13.9	2.4	5.5	2.0	14.1	77.7	94.7	92.0	80.5	73.0	74.3	56.1
bioR single + fullname	40.9	61.3	44.8	36.6	47.9	37.4	19.9	82.4	95.8	93.2	82.3	83.8	80.0	63.9
bioR multi3	88.5	99.5	95.0	88.2	95.0	86.1	71.7	93.9	99.9	96.9	90.0	97.2	93.0	88.8
bioR multi3 + fullname	92.4	99.3	96.9	91.7	97.2	91.9	81.6	93.2	99.9	97.1	90.3	97.5	92.4	85.8
bioR multi5	93.7	100	96.3	91.0	96.8	91.9	89.1	94.8	100	97.8	90.6	97.5	92.8	93.1
bioR multi5 + fullname	94.2	99.6	97.4	90.5	97.3	94.2	90.9	94.5	99.9	96.7	90.9	98.0	93.0	91.5
	QA mean acc	QA b. date	QA b. city	QA univ	QA major	QA C. name	QA C. city	Mix mean acc	Mix b. date	Mix b. city	Mix univ	Mix major	Mix C. name	Mix C. city

Figure 12: Comparison of BIO pretraining + QA finetuning (left) versus their mixed training counterparts (right) under various knowledge augmentations. This is analogous to Figure 3 but for the bioR data family.

We perform experiments on both bioS and bioR data families (refer to Figure 13 and Figure 14), evaluating the P-probing accuracy of first-token and whole-attribute predictions. These figures also compare rank-2 and rank-4 updates on the embedding layer, demonstrating that a large modification to this layer is not crucial for P-probing attribute values.

## H DETAILS ON Q-PROBING

Recall that in Q-probing, we freeze the pretrained GPT model and append a small set of trainable parameters on top for probing purposes. Using GPT2 small as an example, we add:

- a trainable rank- $r$  update on the embedding layer with dimensions of  $50256 \times r$  and  $r \times 768$ ,
- a trainable linear layer with dimensions of  $768 \times M$  for each prediction task that is an  $M$ -class classification problem,
- a batch normalization layer before the linear layer, with trainable affine parameters.

We consider an input sentence that *only* contains a person’s full name, preceded by a starting token and followed by an ending token. After applying all 12 layers of GPT, we extract the hidden states from the last layer at the ending token. For instance, in the GPT2-small model, this is a 768-dimensional vector. We then apply a linear classifier on top to predict the person’s attributes. Similar to P-probing, we adopt a separate set of trainable parameters for each of the 11 classification tasks.

We employ the AdamW optimizer with  $\epsilon = 10^{-6}$ , a weight decay of 0.3, an initial learning rate of 0.001, no warmup, and a linear learning rate decay schedule (reducing to 0 by the end). The batch size is set to 200, and we run a total of 30,000 training steps. During training, we allow the frozen GPT model to use dropout.

Experiments are conducted on both the bioS and the bioR data families, as shown in Figure 15, for first-token prediction and whole-attribute prediction. We compare rank-16 versus rank-64 updates on the embedding layer for the bioS data (or rank-32 versus rank-128 updates for the bioR data). This demonstrates that for Q-probing, a larger modification to the embedding layer is not necessary to probe the desired attribute values.

## I DETAILS ON CELEBRITY AUGEMENTATION

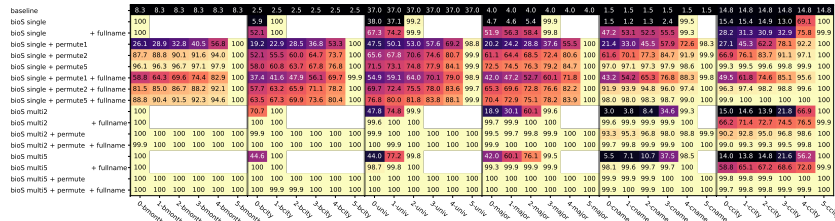
Recall that in the celebrity knowledge augmentation, we introduced an additional set of  $N = 100,000$  individuals and designated them as the celebrity group,  $\mathcal{P}_{\text{cel}}$ . In contrast, the original  $N$  individuals represent the minority group,  $\mathcal{P}_{\text{min}}$ . There is no overlap between these two sets of individuals; specifically, they have distinct full names.

In the main body of this paper (specifically in Figure 4), we considered two choices:

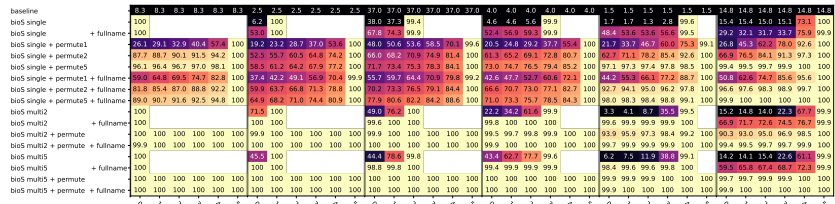
- The minority uses bioS single+permute1, and the celebrity uses bioS multi5+permute. We denote this combination as bioS single+permute1+CEL and compare it to bioS single+permute1.
- The minority uses bioR single, and the celebrity uses bioR multi5. We denote this combination as bioR single+CEL and compare it to bioR single.

(We also compare the latter to bioR single+wiki. By this, we mean that during BIO pretraining, half of the training sentences come from the WikiBook dataset, while the other half come from

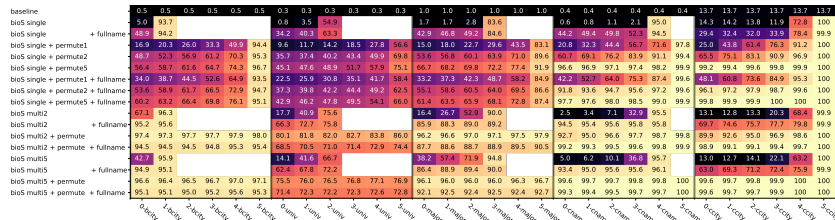




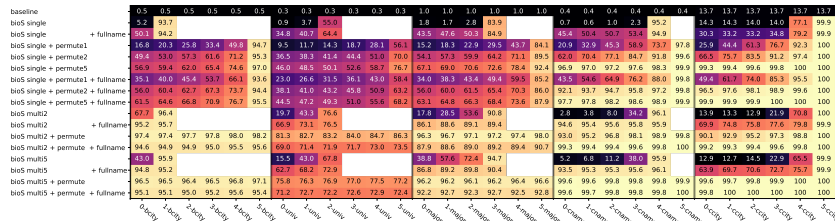
(a) P-probing first-token prediction accuracy; LoRA embedding layer rank = 2



(b) P-probing first-token prediction accuracy; LoRA embedding layer rank = 4



(c) P-probing whole-attribute prediction accuracy; LoRA embedding layer rank = 2



(d) P-probing whole-attribute prediction accuracy; LoRA embedding layer rank = 4

Figure 13: P-probing accuracies on the bioS data (extension of Figure 6). Each row represents a different pretrained model using its associated knowledge augmentation on the bioS data. For every  $i \in \{0, 1, \dots, 5\}$  and  $field \in \{\text{bmonth}, \text{bcity}, \dots\}$ , the column labeled “ $i$ -field” shows the accuracy when predicting the first token / whole attribute of  $field$  from the special position  $i$ .

**Observation.** Comparison between LoRA rank 2 and 4 shows that a rank-2 update on the embedding layer is sufficient for P-probing purposes. The P-probing results for the whole-attribute scenario largely align, but when predicting longer attributes, like “university”, the classification accuracy falls short of 100%. This outcome is consistent with expectations: extracting partial knowledge from subsequent tokens in a lengthy attribute can be difficult, as further detailed in our companion paper Anonymous (2023).





baseline	2.7	0.0	0.5	0.3	1.0	0.4	13.7	8.3	2.5	37.0	4.0	1.5	14.8	0.5	0.3	1.0	0.4	13.7
bioS single	9.7	31.5	6.3	2.3	4.0	1.1	13.8	63.4	1.9	37.5	3.1	0.2	13.1	1.1	0.3	1.4	0.1	11.6
bioS single + fullname	48.9	56.2	58.8	63.0	55.7	50.5	14.1	78.8	47.5	65.4	51.0	47.1	28.9	43.9	31.2	40.8	44.4	28.9
bioS single + permute1	4.4	0.5	3.3	2.4	5.0	3.5	13.7	10.2	1.2	37.5	2.9	0.7	12.8	0.5	0.4	1.5	0.6	12.0
bioS single + permute2	55.2	57.3	48.3	53.1	55.0	51.8	58.8	85.4	42.1	60.1	55.7	51.8	59.1	39.6	30.9	48.2	33.3	58.8
bioS single + permute5	70.0	56.4	57.7	58.3	64.9	90.5	97.7	95.3	97.7	66.6	68.6	96.5	99.1	47.7	40.7	63.0	96.2	98.9
bioS single + permute1 + fullname	31.7	26.6	29.3	36.9	31.1	31.4	37.9	50.1	20.9	42.6	27.1	27.3	34.7	19.3	13.5	21.2	26.8	33.4
bioS single + permute2 + fullname	73.1	69.0	60.6	64.2	64.0	87.9	95.0	77.8	51.0	63.3	61.4	92.1	96.7	48.1	35.0	51.5	91.9	96.6
bioS single + permute5 + fullname	80.2	83.7	67.8	72.6	69.1	93.0	98.6	86.5	59.0	73.5	64.2	98.0	99.9	56.8	40.8	56.6	97.9	99.9
bioS multi2	41.0	100	71.7	93.1	26.1	5.2	14.0	100	70.3	45.1	19.6	0.7	13.0	66.8	14.9	14.8	0.7	11.8
bioS multi2 + fullname	84.0	100	97.7	89.5	97.6	91.3	95.5	100	100	99.7	99.7	99.8	99.2	94.5	82.2	84.2	95.2	72.5
bioS multi2 + permute	91.2	99.3	98.7	89.8	96.7	83.3	83.5	99.9	99.9	99.6	99.4	93.9	90.2	98.2	78.6	95.5	93.1	89.4
bioS multi2 + permute + fullname	96.1	100	98.8	91.3	98.1	93.7	97.8	99.7	100	100	99.9	99.9	99.4	95.4	84.6	88.1	99.4	99.1
bioS multi5	41.0	100	80.8	30.9	43.5	10.2	13.8	100	85.8	40.3	10.1	3.2	13.2	38.1	9.7	36.1	2.3	11.9
bioS multi5 + fullname	82.4	100	98.6	88.4	96.1	91.0	26.8	100	100	98.3	99.4	98.5	93.1	95.2	58.7	87.1	194.1	66.9
bioS multi5 + permute	96.6	100	99.0	91.3	97.7	95.1	98.7	100	100	99.9	100	99.9	99.8	96.1	72.6	94.9	99.6	99.7
bioS multi5 + permute + fullname	96.2	100	98.7	90.6	97.9	93.7	99.0	100	100	99.8	100	100	99.8	95.4	68.1	92.0	99.9	99.8

(a) Q-probing for the bioS data family; LoRA embedding layer rank = 16

baseline	2.7	0.0	0.5	0.3	1.0	0.4	13.7	8.3	2.5	37.0	4.0	1.5	14.8	0.5	0.3	1.0	0.4	13.7
bioS single	9.7	31.5	6.3	2.3	4.0	1.1	13.8	59.5	1.2	37.2	2.4	0.3	9.4	0.9	0.1	1.4	0.3	8.8
bioS single + fullname	48.9	56.2	58.8	63.0	55.7	50.5	14.1	74.9	45.3	64.6	48.1	43.0	26.1	42.7	29.1	38.3	40.9	26.0
bioS single + permute1	4.4	0.5	3.3	2.4	5.0	3.5	13.7	7.7	0.6	37.1	2.6	0.4	10.1	0.5	0.3	1.0	0.4	9.5
bioS single + permute2	55.2	57.3	48.3	53.1	55.0	51.8	58.8	82.7	47.4	57.4	52.5	51.2	54.6	37.2	27.5	44.7	50.2	53.8
bioS single + permute5	70.0	56.4	57.7	58.3	64.9	90.5	97.7	94.3	47.0	63.9	64.5	96.5	99.1	45.1	36.1	59.7	96.4	99.0
bioS single + permute1 + fullname	31.7	26.6	29.3	36.9	31.1	31.4	37.9	43.1	19.0	42.4	22.8	25.4	31.7	17.1	12.1	17.6	24.6	29.7
bioS single + permute2 + fullname	73.1	69.0	60.6	64.2	64.0	87.9	95.0	75.0	47.9	61.6	57.3	91.2	96.4	45.8	32.0	47.9	90.8	96.4
bioS single + permute5 + fullname	80.2	83.7	67.8	72.6	69.1	93.0	98.6	84.4	56.2	71.1	61.8	97.9	99.8	54.1	38.0	53.8	97.8	99.8
bioS multi2	41.0	100	71.7	93.1	26.1	5.2	14.0	100	87.8	41.9	17.6	0.5	10.2	64.6	12.5	12.7	0.4	8.6
bioS multi2 + fullname	84.0	100	97.7	89.5	97.6	91.3	95.5	100	100	99.5	99.8	99.7	92.6	94.6	82.4	84.6	95.2	71.6
bioS multi2 + permute	91.2	99.3	98.7	89.8	96.7	83.3	83.5	99.9	99.9	99.6	99.4	93.7	89.0	98.2	78.9	95.6	92.9	89.0
bioS multi2 + permute + fullname	96.1	100	98.8	91.3	98.1	93.7	97.8	99.7	100	99.8	99.8	99.9	99.4	95.2	83.4	87.9	99.4	99.1
bioS multi5	41.0	100	80.8	30.9	43.5	10.2	13.8	100	86.2	38.9	17.6	3.1	10.9	34.7	7.7	32.7	2.6	9.7
bioS multi5 + fullname	82.4	100	98.6	88.4	96.1	91.0	26.8	100	100	97.8	99.5	98.2	91.9	95.5	59.4	86.8	93.6	65.5
bioS multi5 + permute	96.6	100	99.0	91.3	97.7	95.1	98.7	100	100	99.9	100	99.9	99.7	96.2	73.4	94.8	99.5	99.7
bioS multi5 + permute + fullname	96.2	100	98.7	90.6	97.9	93.7	99.0	100	100	99.7	100	100	99.8	95.2	67.3	92.2	99.9	99.8

(b) Q-probing for the bioS data family; LoRA embedding layer rank = 64

baseline	2.7	0.0	0.5	0.3	1.0	0.4	13.7	8.3	2.5	37.0	4.0	1.5	14.8	0.5	0.3	1.0	0.4	13.7
bioR single	10.0	25.1	13.9	2.4	5.5	2.0	14.1	28.0	1.4	35.7	5.4	0.6	10.5	1.1	0.2	2.7	0.4	9.5
bioR single + fullname	40.8	41.3	44.1	74.4	42.0	27.0	19.9	63.9	28.4	44.9	50.0	24.8	20.7	26.0	16.7	16.2	23.9	20.0
bioR multi3	88.5	99.5	95.0	88.2	95.0	86.1	71.7	99.8	98.3	96.8	98.5	93.0	83.1	96.4	85.4	96.7	92.3	84.3
bioR multi3 + fullname	92.4	99.3	96.9	91.7	97.2	91.9	81.6	99.9	99.7	99.3	99.9	99.6	93.4	97.7	87.0	96.8	99.4	94.5
bioR multi5	93.7	100	96.3	91.0	96.8	91.9	89.1	99.8	99.6	99.2	99.8	99.6	95.6	98.4	88.0	98.9	99.6	96.4
bioR multi5 + fullname	94.2	99.6	97.4	90.5	97.3	94.2	90.9	99.9	99.9	99.6	100	100	97.4	98.3	87.9	98.8	99.9	97.8

(c) Q-probing for the bioR data family; LoRA embedding layer rank = 32

baseline	2.7	0.0	0.5	0.3	1.0	0.4	13.7	8.3	2.5	37.0	4.0	1.5	14.8	0.5	0.3	1.0	0.4	13.7
bioR single	10.0	25.1	13.9	2.4	5.5	2.0	14.1	8.7	1.0	32.2	2.6	0.6	5.7	0.8	0.7	0.9	0.4	5.5
bioR single + fullname	40.8	41.3	44.1	74.4	42.0	27.0	19.9	9.4	24.4	40.9	45.1	20.0	17.5	22.0	13.0	13.5	18.8	16.5
bioR multi3	88.5	99.5	95.0	88.2	95.0	86.1	71.7	99.9	98.8	97.9	99.0	93.8	83.6	96.7	88.7	97.0	93.2	84.7
bioR multi3 + fullname	92.4	99.3	96.9	91.7	97.2	91.9	81.6	99.9	99.7	99.2	99.8	99.6	91.6	97.8	88.6	96.2	99.2	92.9
bioR multi5	93.7	100	96.3	91.0	96.8	91.9	89.1	99.8	99.6	99.4	99.8	99.8	95.5	98.8	89.1	98.8	99.7	96.2
bioR multi5 + fullname	94.2	99.6	97.4	90.5	97.3	94.2	90.9	99.9	99.8	99.7	100	100	96.8	98.4	87.2	98.8	99.8	97.2

(d) Q-probing for the bioR data family; LoRA embedding layer rank = 128

Figure 15: Q-probing accuracies (extension of Figure 7). Each row denotes a pretrained model with its specific knowledge augmentation. The left block reiterates QA finetune accuracies from Figure 3 and Figure 12. The middle showcases Q-probing accuracies on the first-token prediction for the six attributes, and the right focuses on Q-probing for the “whole-attribute” prediction.

**Observation.** Comparison between LoRA ranks show that a rank-16 (resp. rank-32) update on the embedding layer is sufficient for Q-probing purposes on bioS (resp. bioR). Q-probing results on the bioR data family closely mirror those on bioS. Incorporating additional knowledge augmentations in the pretrain data enables the Q-probing accuracies to significantly improve.

the bioR single data.)<sup>21</sup>

Note that in both cases, each individual in the minority group has only one biographical entry, while each individual in the celebrity group has five biographical entries. Thus, during BIO pretraining, the BIO data on  $\mathcal{P}_{\text{cel}}$  appear with a  $1/6$  chance.

In this appendix, we explore a broader set of augmentation options.

- The minority uses bioS single and the celebrity uses bioS multi5+permute, denoted as bioS single+CEL. We compare this to bioS single. In this scenario, the celebrity and minority groups **have biographical entries in different formats**: the entries of the celebrity group are *randomly shuffled*, while those of the minority group follow a *fixed order* (see (2.1)). The QA test accuracy on the minority group increases with the addition of the celebrity group, but not to the same extent as in the bioS single+permute1+CEL case.
- The minority uses bioS single+permute1+fullname and the celebrity uses bioS multi5+permute, denoted as bioS single+permute1+fullname+CEL. We compare this to bioS single+permute1+fullname. In this scenario, the celebrity and minority groups **have their biographical entries in different formats**: the minority group uses the fullname augmentation, repeating the individual’s full name in each sentence, while the celebrity group only mentions the fullname once. The QA test accuracy on the minority group increases with the assistance of the celebrity group, but not as much as in the bioS single+permute1+CEL case.
- The minority uses bioR single+fullname and the celebrity uses bioR multi5+fullname, denoted as bioR single+fullname+CEL. We compare this to bioR single+fullname. In this case, the celebrity and minority groups **have their biographical entries in the same format**, leading to a significant increase in QA test accuracy to 82.2%.

(We also compare this to bioR single+fullname+wiki, where during BIO pretraining, half of the training sentences come from the WikiBook dataset, and the other half from the bioR single+fullname data. C.f. Remark 4.1)

The transformer model is pretrained on the combined set of biographies  $\mathcal{P}_{\text{cel}} \cup \mathcal{P}_{\text{min}}$  and then fine-tuned using QAs from the celebrity group  $\mathcal{P}_{\text{cel}}$ . We evaluate the model’s QA generation accuracy on the  $\mathcal{P}_{\text{min}}$  group.<sup>22</sup> Our findings are reported in Figure 16.

baseline	2.7	0.0	0.5	0.3	1.0	0.4	13.7
bioS single + permute1	4.4	0.5	3.3	2.4	5.0	3.5	13.7
bioS single + permute1 + CEL	86.8	98.3	96.8	90.7	90.2	71.7	80.1
bioS single	9.7	33.5	6.3	2.3	4.0	1.1	13.8
bioS single + CEL	63.2	95.1	91.4	70.6	60.6	45.0	23.1
bioS single + permute1 + fullname	31.7	26.6	29.3	36.9	31.1	31.4	37.9
bioS single + permute1 + fullname + CEL	52.7	62.3	67.6	60.4	55.6	33.1	39.0
bioR single	10.0	25.1	13.9	2.4	5.5	2.0	14.1
bioR single + CEL	7.3	18.4	5.2	2.6	4.3	1.8	14.1
bioR single + wiki	76.3	94.3	85.3	82.9	79.4	67.0	56.6
bioR single + fullname	40.9	61.3	44.8	36.6	47.9	37.4	19.9
bioR single + fullname + wiki	27.7	39.3	33.3	25.8	33.5	22.2	14.6
bioR single + fullname + CEL	82.2	97.2	89.7	87.3	88.9	77.8	60.2

QA mean acc    QA b\_date    QA b\_city    QA univ    QA major    QA c\_name    QA c\_city

Figure 16: QA finetune accuracy on the *minority group* with versus without celebrity data in the pretraining process. This is an extension to Figure 4, and the details are given in Appendix I.

**Observation.** The augmentation effect from the celebrity data may be weakened if the minority group uses differently formatted BIO data, such as using full names when the celebrity does not (see bioS single+permute1+fullname+CEL), or maintaining a fixed sentence order when the celebrity does not (see bioS single+CEL). We also conducted an experiment where both the celebrity and minority used bioR data with full name augmentation. In all cases, incorporating celebrity data significantly improved QA test accuracy for the minority group.

<sup>21</sup>Recall that BERT and RoBERTa were trained on a combination of BookCorpus (Zhu et al., 2015) and English Wikipedia, which totals 16GB of uncompressed text (Kenton & Toutanova, 2019; Liu et al., 2019). We use this same 16GB WikiBook dataset.

<sup>22</sup>We also considered other fine-tuning variations, such as QA finetuning with half of  $\mathcal{P}_{\text{min}}$  as training and half as testing, but found negligible differences.

**P-probing and Q-probing.** We incorporate P-probing and Q-probing results for our celebrity case. The inclusion of celebrity data enhances the model’s structural knowledge storage, *even for minority groups*. Figure 17 demonstrates that knowledge about minority groups is often stored in earlier tokens. This confirms that for *minority groups*, individual full names can more directly encode the six target attributes, due to the introduction of celebrity data. This accounts for the high knowledge-extraction QA accuracies.

baseline	8.3	8.3	8.3	8.3	8.3	2.5	2.5	2.5	2.5	2.5	37.0	37.0	37.0	37.0	37.0	4.0	4.0	4.0	4.0	4.0	1.5	1.5	1.5	1.5	1.5	14.8	14.8	14.8	14.8	14.8	
bioS single	100	100	100	100	100	5.9	100	100	100	100	38.0	37.1	99.2	100	100	4.7	4.6	5.4	99.9	100	1.5	1.2	1.3	2.4	99.5	15.4	15.4	14.9	13.0	69.1	
bioS single + permute1	26.1	28.9	32.8	40.5	56.8	100	19.2	22.9	28.5	36.8	100	47.5	50.1	53.0	57.6	99.2	100	20.2	24.2	28.8	37.6	100	21.4	33.0	45.5	57.9	72.6	98.3	27.1	45.3	62.2
bioS single + permute1 + CEL	96.6	98.6	98.8	98.8	98.9	99.1	95.4	96.1	96.3	96.7	97.1	98.6	94.3	95.2	95.5	95.8	96.6	97.6	93.6	94.5	95.1	95.6	96.7	98.6	78.2	81.6	85.4	88.2	92.0	96.0	
bioR single	95.5	96.4	96.9	96.1	93.0	100	97.4	95.4	95.1	94.9	94.8	100	43.5	45.5	46.6	73.1	79.9	98.4	33.1	37.4	49.3	73.1	67.7	98.7	18.2	11.3	22.5	18.5	64.6	99.2	
bioR single + wiki	50.7	35.3	38.8	35.2	53.4	99.9	38.6	34.4	35.7	43.6	55.4	100	43.6	45.5	46.8	72.6	78.2	98.9	33.4	37.1	44.2	72.6	67.2	98.7	17.6	10.8	21.7	17.7	33.1	99.4	
bioR single + CEL	99.4	99.7	99.7	99.7	99.6	99.7	91.6	94.9	92.2	93.6	94.1	97.4	94.5	95.9	97.4	97.7	98.2	99.1	93.2	95.2	96.6	97.4	97.6	98.7	81.2	83.4	86.5	87.8	93.3	97.8	
bioR single + fullname	77.0	70.7	70.7	71.0	76.0	100	53.7	67.7	56.6	62.3	73.6	99.2	51.7	54.3	72.1	78.4	84.1	98.6	64.0	66.7	74.8	85.0	85.7	98.7	42.2	40.2	48.3	46.3	87.1	99.1	
bioR single + fullname + wiki	74.6	69.2	72.9	72.8	78.9	99.7	54.5	66.4	54.9	60.5	74.8	98.2	49.4	52.9	72.5	77.7	84.6	97.3	56.8	61.4	72.4	83.6	85.0	98.7	34.4	32.4	43.2	42.8	84.6	98.8	
bioR single + fullname + CEL	98.9	99.3	99.5	99.3	99.2	99.3	93.9	96.0	94.9	95.5	95.7	97.6	93.9	95.2	97.3	97.7	97.8	99.1	93.9	95.4	97.5	98.0	98.3	99.8	86.7	87.5	90.6	90.9	94.7	98.5	
0.00																															

(a) P-probing first-token prediction accuracy; LoRA embedding layer rank = 2

baseline	0.5	0.5	0.5	0.5	0.5	0.3	0.3	0.3	0.3	0.3	1.0	1.0	1.0	1.0	1.0	0.4	0.4	0.4	0.4	0.4	13.7	13.7	13.7	13.7	13.7				
bioS single	2.5	23.7	26.0	31.3	48.8	94.4	96	11.7	14.2	18.5	27.8	96.9	15.0	18.0	22.7	29.6	43.5	83.1	98.8	99.3	99.4	99.7	116	97.8	95.0	83.8	81.4	76.3	72.8
bioS single + permute1	36.9	20.5	26.0	31.3	48.8	94.4	96	11.7	14.2	18.5	27.8	96.9	15.0	18.0	22.7	29.6	43.5	83.1	98.8	99.3	99.4	99.7	116	97.8	95.0	83.8	81.4	76.3	72.8
bioS single + permute1 + CEL	89.7	90.2	90.5	90.7	91.2	92.7	93.9	94.4	95.3	95.4	95.8	97.3	93.1	84.3	84.8	86.2	86.2	86.7	76.0	80.0	84.0	87.6	91.5	95.7	78.3	78.0	86.5	91.2	96.3
bioR single	34.9	52.5	33.8	42.6	51.7	96.7	8.5	11.2	33.0	40.0	46.2	65.4	28.5	32.3	42.9	63.4	60.1	87.9	17.2	10.5	21.6	17.7	81.4	98.6	21.8	28.7	30.1	36.2	68.7
bioR single + wiki	31.4	51.3	33.1	40.5	51.7	93.8	7.7	10.0	27.2	33.2	39.3	56.6	28.8	32.3	42.0	62.5	56.8	84.9	16.7	10.4	20.7	17.1	78.9	98.4	20.7	28.9	29.7	36.5	65.7
bioR single + CEL	98.7	92.0	89.6	90.6	90.9	93.4	75.5	77.3	80.1	80.9	81.8	83.8	89.3	91.5	92.7	93.7	93.9	95.7	80.0	82.4	84.7	86.3	92.0	96.7	69.8	75.5	76.1	79.9	83.6
bioR single + fullname	93.2	64.3	52.9	58.1	70.3	95.4	25.1	28.0	44.3	51.0	55.8	71.0	54.9	58.7	65.6	75.9	76.6	88.6	31.4	39.0	47.4	45.1	84.2	98.3	32.6	39.3	41.9	47.2	76.8
bioR single + fullname + wiki	93.3	64.8	53.2	56.4	71.6	93.3	26.0	23.4	41.1	46.1	51.1	64.6	58.0	58.4	63.3	73.6	74.3	88.9	32.0	39.2	48.1	48.1	81.7	97.7	27.7	35.1	37.3	43.6	70.5
bioR single + fullname + CEL	98.0	92.3	90.9	92.0	91.7	93.5	84.9	83.5	88.5	79.1	79.2	74.2	90.3	91.9	93.9	94.8	95.2	96.0	85.4	86.2	89.5	89.8	93.2	98.1	73.1	73.2	80.5	83.2	85.4
bioS single + permute1 + CEL	98.0	92.3	90.9	92.0	91.7	93.5	84.9	83.5	88.5	79.1	79.2	74.2	90.3	91.9	93.9	94.8	95.2	96.0	85.4	86.2	89.5	89.8	93.2	98.1	73.1	73.2	80.5	83.2	85.4

(b) P-probing whole-attribute prediction accuracy; LoRA embedding layer rank = 2

Figure 17: P-probing accuracies on the *minority group* with or without **celebrity** data. Each row represents a different pretrained model using its associated knowledge augmentation on the bioS data (*with or without celebrity data*). For every  $i \in \{0, 1, \dots, 5\}$  and  $field \in \{\text{bmonth}, \text{bcity}, \dots\}$ , the column labeled “ $i$ - $field$ ” shows the accuracy when predicting the first token / whole attribute of  $field$  from the special position  $i$ , among individuals in the minority group.

baseline	2.7	0.0	0.5	0.3	1.0	0.4	13.7	8.3	2.5	37.0	4.0	1.5	14.8	0.5	0.3	1.0	0.4	13.7
bioS single	9.7	33.5	6.3	2.3	4.0	1.1	13.8	63.4	1.9	37.5	3.1	0.3	13.1	1.1	0.3	1.4	0.3	11.6
bioS single + permute1	4.4	0.5	3.3	2.4	5.0	3.5	13.7	10.2	1.2	37.5	2.9	0.7	12.8	0.5	0.4	1.5	0.6	12.0
bioS single + permute1 + CEL	86.8	98.3	96.8	90.7	90.2	71.7	80.1	98.2	97.6	96.0	93.5	80.5	80.3	91.7	58.4	85.3	80.2	79.4
bioR single	10.0	25.1	13.9	2.4	5.5	2.0	14.1	28.0	1.4	35.7	5.4	0.6	10.5	1.1	0.7	2.7	0.4	9.5
bioR single + wiki	7.3	18.4	5.2	2.6	4.3	1.8	14.1	11.0	1.0	35.9	5.3	0.6	11.2	0.7	0.3	2.7	0.6	10.4
bioR single + CEL	76.3	94.3	85.3	82.9	79.4	67.0	56.6	97.9	89.8	89.4	84.2	75.2	57.9	87.0	71.5	80.3	73.5	59.1
bioR single + fullname	40.9	61.3	44.8	36.6	47.9	37.4	19.9	65.9	28.4	44.9	50.0	24.8	20.7	26.0	16.7	41.4	23.9	20.0
bioR single + fullname + wiki	27.7	39.3	33.3	25.8	33.5	22.2	14.6	61.1	22.5	40.4	32.8	13.5	14.7	21.4	12.0	25.8	11.8	13.7
bioR single + fullname + CEL	82.2	97.2	89.7	87.3	88.9	77.8	60.2	98.2	95.7	92.2	92.3	87.3	69.6	92.7	70.9	88.8	85.7	71.5

Figure 18: Q-probing accuracies on the *minority group* with or without **celebrity** data. Each row denotes a pretrained model with its specific knowledge augmentation. The left block reiterates QA finetune accuracies on the minority group (same as Figure 16). The middle showcases Q-probing accuracies on the first-token prediction for the six attributes of individuals in the minority group, and the right focuses on Q-probing for the “whole-attribute” prediction. Recall we have used a LoRA embedding rank 16 for the bioS data and rank 32 for the bioR data (see Appendix H).

## J DETAILS ON BERT EXPERIMENT

Recall that GBERT is a bi-directional variant of GPT2, using the same tokenizer, as detailed in Appendix D. It is similar to BERT, but its architecture closely resembles GPT2 for direct comparison. We use GBERT for the following tasks: (1) BIO pretrain, (2) BIO+QA mix training, (3) QA finetune from BIO pretrain, and (4) Q-probing from BIO pretrain. Since we only apply GBERT to the bioS data family to demonstrate a negative result, we utilize the same architecture size as GPT2-small.

For BIO pretrain and BIO+QA mix training, we use the AdamW optimizer with weight decay 0.1,  $\varepsilon = 10^{-6}$ , an initial learning rate of 0.0003, a 1000-step linear warmup, and cosine learning rate decay (from 0.0003 to 0.00003). We use a batch size of 96 for 150000 training steps on the bioS dataset. This is *twice the training time* compared to the 80000 steps used for GPT2 small on the same

dataset, as we are presenting a negative result on GBERT. For BIO+QA mix training, we tested both  $QA_r = 0.2$  and  $QA_r = 0.8$  and report the best test accuracy.

For QA finetune, we tested four LoRA variants and report their best accuracy.<sup>23</sup> We use the AdamW optimizer with weight decay 0.01 and an initial learning rate of 0.0003 for all tasks, with linear learning rate decay (down to 0). We use a batch size of 48 for 50000 training steps.

For Q-probing, we use the AdamW optimizer with  $\varepsilon = 10^{-6}$ , weight decay 0.3, an initial learning rate of 0.001, no warmup, linear learning rate decay (down to 0), a batch size of 200, and 30000 training steps. This is identical to the procedure outlined in Appendix H.

All of our results were presented in the same Figure 9.

---

<sup>23</sup>Specifically, we tested rank-8 or rank-32 update on the query/value matrices, and rank-128 update or full fine-tuning on the embedding layer.