A. Additional Results on Variance and Calibration

Table 2 shows an example of the sensitivity to ordering.

Prompt (test input not shown)	**Acc.**
Review: the whole thing 's fairly lame , making it par for the course for disney sequels . Answer: Negative Review: this quiet , introspective and entertaining independent is worth seeking . Answer: Positive	88.5%
Review: this quiet , introspective and entertaining independent is worth seeking . Answer: Positive Review: the whole thing 's fairly lame , making it par for the course for disney sequels . Answer: Negative	51.3%

Table 2. Top: a prompt consisting of two training examples (the test input is not shown) that leads to good test accuracy for GPT-3 2.7B (88.5%). Bottom: simply *reversing the order* of the two examples causes the accuracy to drop to near random chance (51.3%).

Table 3 demonstrates that the choice of content-free input does affect accuracy, however, many good choices exist.

Content-free Input	**SST-2**	**AGNews**
Uncalibrated Baseline	66.5	48.5
N/A	74.2	64.5
[MASK]	74.5	63.8
‘’	72.9	64.7
N/A, [MASK], ‘’	79.0	66.5
the	69.1	59.0
abc	77.5	57.3
the man.	79.4	62.0
dasjhasjkdhjskdhds	79.3	64.5
nfjkhdvy84tr9bpuirvwe	78.4	65.5

Table 3. We show the accuracy for 1-shot SST-2 and 0-shot AGNews over different choices for the content-free input. The choice of content-free input matters, however, *many good choices exist*. The token ‘’ indicates the empty string. Recall that in our experiments, we ensemble over N/A, [MASK], and the empty string.

Figure 9 shows how GPT-3 accuracy changes as the prompt format is varied for LAMA, with and without calibration.

Table 4 shows the effect of calibration for GPT-2.

B. Prompt Formats Used

Tables 5 and 6 show the default prompt format used for all tasks. Table 7 shows the 15 different formats used when studying the effect of prompt format for SST-2.

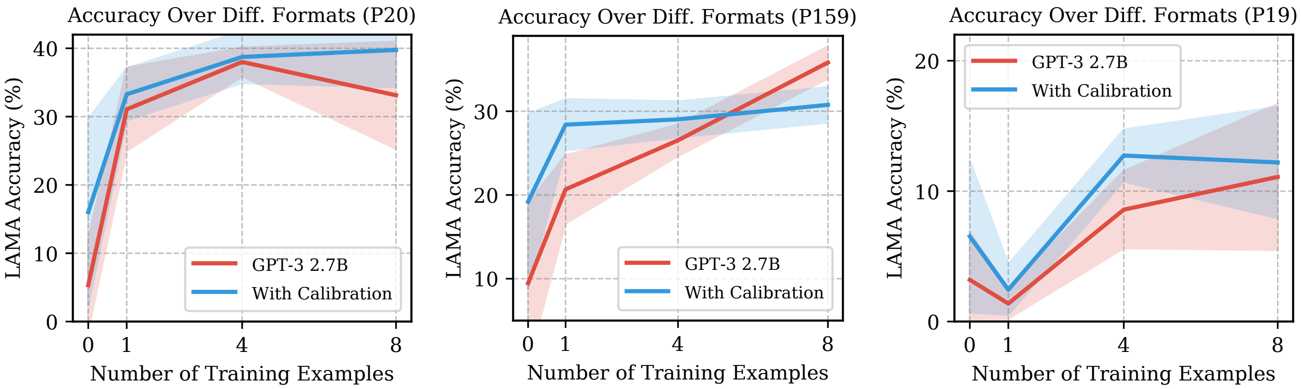

Figure 9. Contextual calibration improves GPT-3's accuracy across various prompt formats for LAMA. We plot GPT-2 2.7B's mean accuracy over 15 different formats for the LAMA "place of death" relation (P20), "Headquarter Location" relation (P159), and "place of birth" relation (P19).

Dataset	**LM**	**0-shot**		**1-shot**		**4-shot**		**8-shot**	
		Baseline	Ours	Baseline	Ours	Baseline	Ours	Baseline	Ours
Text Classification									
AGNews	GPT-2	44.0$_{0.0}$	**60.0**$_{0.0}$	45.4$_{8.4}$	**67.9**$_{5.7}$	44.6$_{12.2}$	**58.0**$_{13.6}$	57.1$_{11.6}$	**63.1**$_{7.3}$
TREC	GPT-2	24.0$_{0.0}$	**37.3**$_{0.0}$	21.5$_{5.2}$	**41.1**$_{2.6}$	23.1$_{5.9}$	**44.2**$_{2.2}$	32.7$_{7.5}$	**44.1**$_{3.6}$
CB	GPT-2	**44.6**$_{0.0}$	17.9$_{0.0}$	**49.6**$_{10.0}$	47.1$_{12.2}$	40.0$_{8.3}$	**55.4**$_{7.3}$	48.9$_{5.7}$	**63.2**$_{1.4}$
RTE	GPT-2	**51.0**$_{0.0}$	48.5$_{0.0}$	**57.6**$_{2.1}$	56.3$_{2.4}$	53.2$_{6.0}$	**57.5**$_{1.8}$	54.9$_{3.0}$	**57.7**$_{1.29}$
SST-2	GPT-2	60.0$_{0.0}$	**82.0**$_{0.0}$	66.7$_{17.9}$	**73.0**$_{11.4}$	64.9$_{8.4}$	**73.8**$_{10.9}$	54.5$_{4.6}$	**64.6**$_{8.8}$
DBPedia	GPT-2	**64.3**$_{0.0}$	58.3$_{0.0}$	33.6$_{18.9}$	**69.5**$_{9.4}$	53.0$_{14.8}$	**75.3**$_{8.1}$	66.0$_{3.6}$	**74.3**$_{8.7}$
Fact Retrieval									
LAMA	GPT-2	14.0$_{0.0}$	**22.7**$_{0.0}$	29.7$_{1.8}$	**31.6**$_{1.3}$	35.8$_{3.8}$	**37.4**$_{3.4}$	**42.5**$_{1.3}$	**42.5**$_{1.4}$
Information Extraction									
MIT-G	GPT-2	7.7$_{0.0}$	**10.0**$_{0.0}$	32.9$_{10.0}$	**41.2**$_{4.1}$	44.3$_{6.5}$	**47.7**$_{5.8}$	56.9$_{2.5}$	**59.5**$_{2.5}$
MIT-D	GPT-2	29.3$_{0.0}$	**41.7**$_{0.0}$	26.2$_{10.5}$	**58.8**$_{4.8}$	70.5$_{2.5}$	**75.4**$_{1.8}$	77.1$_{4.4}$	**78.1**$_{3.9}$
ATIS-A	GPT-2	15.1$_{0.0}$	**35.5**$_{0.0}$	41.5$_{11.7}$	**51.4**$_{7.5}$	55.1$_{18.9}$	**65.8**$_{11.7}$	63.4$_{10.6}$	**69.9**$_{10.4}$
ATIS-D	GPT-2	1.0$_{0.0}$	**2.5**$_{0.0}$	62.3$_{9.2}$	**68.7**$_{4.3}$	81.1$_{3.6}$	**83.2**$_{7.2}$	81.8$_{4.5}$	**83.9**$_{5.0}$

Table 4. **Contextual calibration improves accuracy for GPT-2.** This table is analogous to Table 1 but shows results for GPT-2 XL.

Task	Prompt	Label Names
SST-2	Review: This movie is amazing! Sentiment: Positive Review: Horrific movie, don't see it. Sentiment:	Positive, Negative
AGNews	Article: USATODAY.com - Retail sales bounced back a bit in July, and new claims for jobless benefits fell last week, the government said Thursday, indicating the economy is improving from a midsummer slump. Answer: Business Article: New hard-drive based devices feature color screens, support for WMP 10. Answer:	World, Sports, Business, Technology
TREC	Classify the questions based on whether their answer type is a Number, Location, Person, Description, Entity, or Abbreviation. Question: How did serfdom develop in and then leave Russia? Answer Type: Description Question: When was Ozzy Osbourne born? Answer Type:	Number, Location, Person, Description, Entity, Abbreviation
DBPedia	Classify the documents based on whether they are about a Company, School, Artist, Athlete, Politician, Transportation, Building, Nature, Village, Animal, Plant, Album, Film, or Book. Article: Geoffrey D. Falksen (born July 31 1982) is an American steampunk writer. Answer: Artist Article: The Perrin River is a 1.3-mile-long (2.1 km) tidal river in the U.S. state of Virginia. It is a small inlet on the north shore of the York River near that river's mouth at Chesapeake Bay. Answer:	Company, School, Artist, Athlete, Politician, Transportation, Building, Nature, Village, Animal, Plant, Album, Film, Book
CB	But he ended up eating it himself. I was reluctant to kiss my mother, afraid that somehow her weakness and unhappiness would infect me. Naturally I didn't think for a minute that my life and spirit could stimulate her. question: her life and spirit could stimulate her mother. True, False, or Neither? answer: Neither Valence the void-brain, Valence the virtuous valet. Why couldn't the figger choose his own portion of titanic anatomy to shaft? Did he think he was helping? question: Valence was helping. True, False, or Neither? answer:	True, False, Neither
RTE	Others argue that Mr. Sharon should have negotiated the Gaza pullout - both to obtain at least some written promises of better Palestinian behavior, and to provide Mr. Abbas with a prime prize to show his people that diplomacy, not violence, delivered Gaza. question: Mr. Abbas is a member of the Palestinian family. True or False? answer: False The program will include Falla's ”Night in the Gardens of Spain,” Ravel's Piano Concerto in G, Berlioz's Overture to ”Beatrice and Benedict,” and Roy Harris' Symphony No. 3. question: Beatrice and Benedict is an overture by Berlioz. True or False? answer:	True, False

Table 5. The prompts used for text classification. We show one training example per task for illustration purposes. The right column shows the label names (to make predictions, we check the LM's probability for these tokens).

Task	Prompt
LAMA	Alexander Berntsson was born in Sweden
	Khalid Karami was born in
ATIS (Airline)	Sentence: what are the two american airlines flights that leave from dallas to san francisco in the evening Airline name: american airlines
	Sentence: list a flight on american airlines from toronto to san diego Airline name:
ATIS (Depart Date)	Sentence: please list any flight available leaving oakland california tuesday arriving philadelphia wednesday Depart date - Day name: tuesday
	Sentence: show me all all flights from pittsburgh to atlanta on wednesday which leave before noon and serve breakfast Depart date - Day name:
MIT Movies (Genre)	Sentence: last to a famous series of animated movies about a big green ogre and his donkey and cat friends Genre: animated
	Sentence: what is a great comedy featuring the talents of steve carell as a loser looking for a friend Genre:
MIT Movies (Director)	Sentence: in 2005 director christopher nolan rebooted a legendary dc comics superhero with a darker grittier edge in which movie Director: christopher nolan
	Sentence: what 1967 mike nichols film features dustin hoffman in romantic interludes with anne bancroft as mrs robinson Director:

Table 6. The prompts used for generation tasks. We show one training example per task for illustration purposes.

Format ID	Prompt	Label Names
1	Review: This movie is amazing! Answer: Positive Review: Horrific movie, don’t see it. Answer:	Positive, Negative
2	Review: This movie is amazing! Answer: good Review: Horrific movie, don’t see it. Answer:	good, bad
3	My review for last night’s film: This movie is amazing! The critics agreed that this movie was good My review for last night’s film: Horrific movie, don’t see it. The critics agreed that this movie was	good, bad
4	Here is what our critics think for this month’s films. One of our critics wrote ”This movie is amazing!”. Her sentiment towards the film was positive. One of our critics wrote ”Horrific movie, don’t see it”. Her sentiment towards the film was	positive, negative
5	Critical reception [edit] In a contemporary review, Roger Ebert wrote ”This movie is amazing!”. Entertainment Weekly agreed, and the overall critical reception of the film was good. In a contemporary review, Roger Ebert wrote ”Horrific movie, don’t see it”. Entertainment Weekly agreed, and the overall critical reception of the film was	good, bad
6	Review: This movie is amazing! Positive Review? Yes Review: Horrific movie, don’t see it. Positive Review?	Yes, No
7	Review: This movie is amazing! Question: Is the sentiment of the above review Positive or Negative? Answer: Positive Review: This movie is amazing! Question: Is the sentiment of the above review Positive or Negative? Answer:	Positive, Negative
8	Review: This movie is amazing! Question: Did the author think that the movie was good or bad? Answer: good Review: This movie is amazing! Question: Did the author think that the movie was good or bad? Answer:	good, bad
9	Question: Did the author of the following tweet think that the movie was good or bad? Tweet: This movie is amazing! Answer: good Question: Did the author of the following tweet think that the movie was good or bad? Tweet: Horrific movie, don’t see it Answer:	good, bad
10	This movie is amazing! My overall feeling was that the movie was good Horrific movie, don’t see it. My overall feeling was that the movie was	good, bad
11	This movie is amazing! I liked the movie. Horrific movie, don’t see it. I	liked, hated
12	This movie is amazing! My friend asked me if I would give the movie 0 or 5 stars, I said 5 Horrific movie, don’t see it. My friend asked me if I would give the movie 0 or 5 stars, I said	0, 5
13	Input: This movie is amazing! Sentiment: Positive Input: Horrific movie, don’t see it. Sentiment:	Positive, Negative
14	Review: This movie is amazing! Positive: True Review: Horrific movie, don’t see it. Positive:	True, False
15	Review: This movie is amazing! Stars: 5 Review: Horrific movie, don’t see it. Stars:	5, 0

Table 7. The different prompt formats used when studying the effect of format for SST-2. We show one training example for illustration.