Exploring Versatile Generative Language Model Via Parameter-Efficient Transfer Learning

Zhaojiang Lin[*], **Andrea Madotto**[*], **Pascale Fung**
Center for Artificial Intelligence Research (CAiRE)
Department of Electronic and Computer Engineering
The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
{zlinao,amadotto}@connect.ust.hk,
pascale@ece.ust.hk

A Supplemental Material

A.1 Model details

Figure 1 illustrates a detailed version of VLM. VLM shares a GPT-2 back-bone and for each task, the model looks up a set of task embeddings for modeling different input structures and chooses the corresponding adapter.

A.2 Experiment details

In this section, we will describe the dataset, evaluation metrics, dataset preprocessing and training details for each task.

Conversational Question Answering (CQA) CoQA (Reddy et al., 2019) is a free-form conversational question answering dataset. The task is to answer the questions in a conversation. Each turn in the conversation contains a question, and we need to answer the questions based on conversation histories and documents. We use *document*, *question*, and *answer* segment embedding to help the model to distinguish the document and alternating questions and answers in the input sequence. We fine-tune the full GPT2-small or VLM (trainable adapter with a fixed GPT2-small) for five epochs with the Adam optimizer. For distillation we only fine-tune VLM for three epochs. We set the batch size to 16 and limit the maximum length of the document to 400 tokens and only retain the last two turns of questions and answers in the dialogue history. Following Reddy et al. (2019) we use the F1 score as evaluation metrics.

Summarization (SUM) CNN/Daily-Mail is a benchmark (Hermann et al., 2015; Nallapati et al., 2016) for text summarization. We use *article*, *summary* segment embedding to divide the article and the summary. We fine-tune the full GPT2-small and VLM for 10 epochs with the Adam optimizer. For distillation, we only fine-tune VLM for five epochs. We set the batch size to 32 and limit the maximum length of the article to 400 tokens and that of the summary to 130 tokens. We use the ROUGE-1, ROUGE-2, and ROUGE-L scores (Lin, 2004) as evaluation metrics.

Neural Machine Translation (NMT) We use the spoken German-English translation dataset IWSLT (Cettolo et al., 2016) as our NMT benchmark. We use *source*, *target* segment embedding to divide the source language and the target language. We fine-tune the full GPT2-small, VLM and distillated VLM for 8 epochs with the Adam optimizer. We set the batch size to 32 and limit the maximum length of the source and target sequence to 100 tokens. We use BLEU (Papineni et al., 2002) as the evaluation metric.

Persona Dialogue (DLG) The Persona-Chat dataset (Zhang et al., 2018) is a persona-grounded multi-turn converstion dataset. We use *persona*, *system*, *user* segment embedding to help the model to distinguish the persona, alternating system utterance and user utterance in an input sequence. We fine-tune the full GPT2-small or VLM for three epochs with the Adam optimizer. We set the batch size to 16 and only retain the last five utterances in the dialogue history. We use perplexity, BLEU, and Consistency score (Madotto et al., 2019) as evaluation metrics.

Natural Language Generation (NLG) The natural language generation challenge (Dušek et al., 2019) is a dataset for building a response generation module for task-oriented dialogue systems. Given a set of response attributes, the model needs to generate responses. For example, when the input attribute is *name[The Wrestlers]*, *priceRange[cheap]*, *customerRating[low]*, the output should be *The wrestlers offers competitive prices, but is not highly*

[*] Equal contributions.

rated by customers. We use a set of attribute segment embedding to segment the input attributes. We fine-tune the full GPT2-small and VLM for 10 epochs with the Adam optimizer. We set the batch size to 32 and use BLUE (Papineni et al., 2002) , ROUGE (Lin, 2004), NIST (Lin and Och, 2004), METEOR (Denkowski and Lavie, 2014) and CiDER (Vedantam et al., 2015) as evaluation metrics.

Computational Cost Fine-tuning VLM requires around 80%-90% GPU memory compared to full-finetune the whole GPT-2 model, as it only updates the small ratio of parameters. And both models have similar training cost, we report the training speed with single GTX 1080 Ti:

Task	Training Speed	Training set size
SUM	7.5h/epoch	300, 000
NMT	1.6h/epoch	200, 000
DLG	1.5h/epoch	130, 000
QA	5.0h/epoch	100, 000
NLG	0.2h/epoch	42, 000

A.3 Detailed Results

In this section, we report the detailed results for each task in Tables 2-6. We use a greedy decoding strategy for all the tasks.

A.4 Example

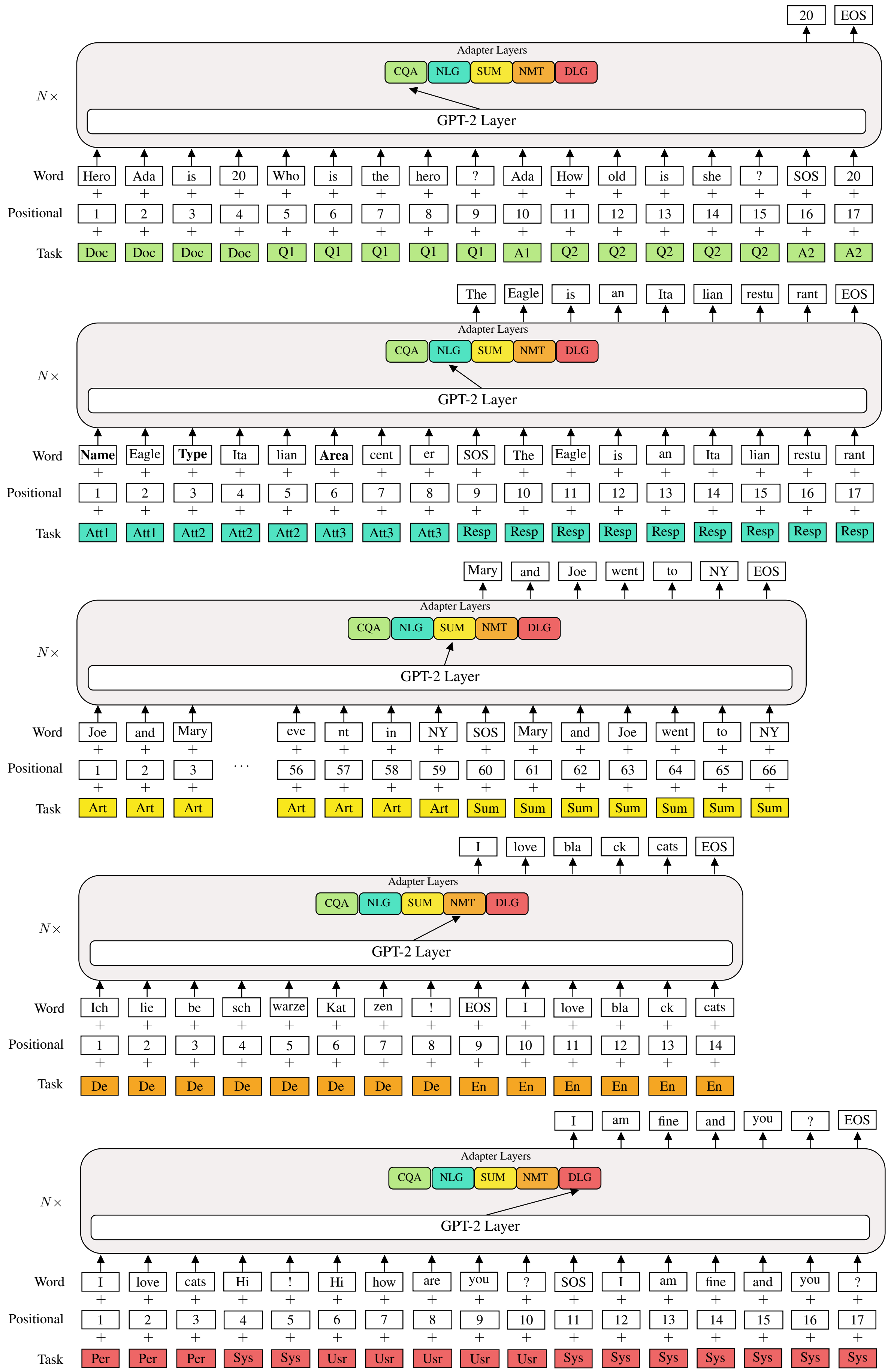

Figure 1: A detailed version of VLM. VLM shares a GPT-2 back-bone and for each task, the model looks up a set of task embeddings and chooses the corresponding adapter.

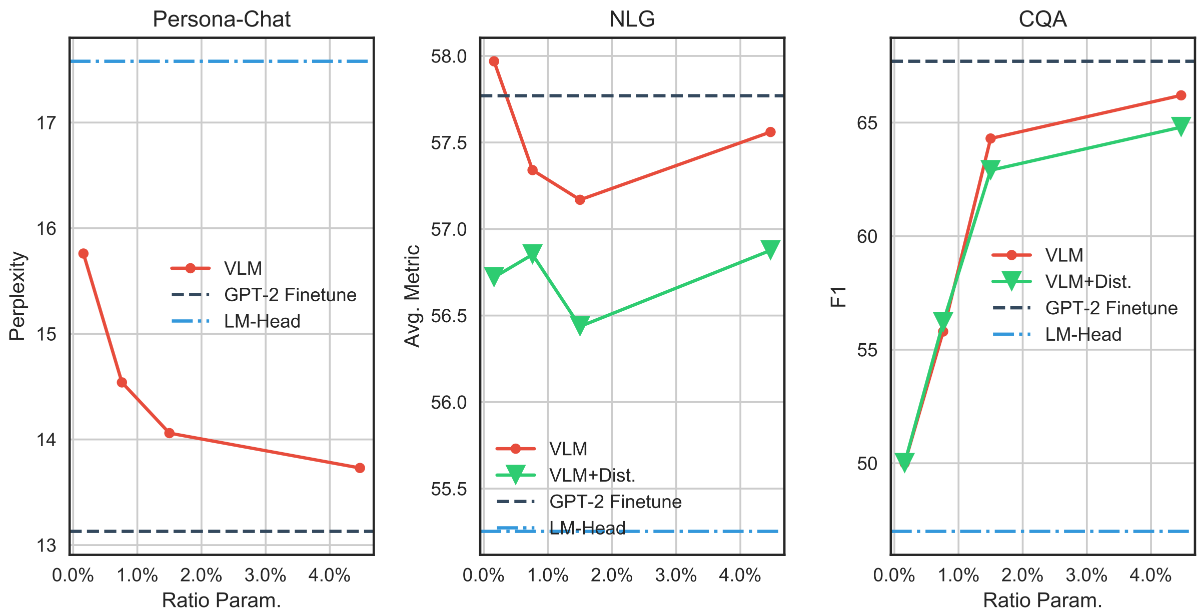

Figure 2: Performance comparison among different ratios of additional parameters. Here we can see that knowledge distillation does not improve the performance of the NLG task because of the small gap between VLM and the full fine-tuned GPT-2. Instead for the dialogue and QA tasks, the gold target is always better than the distillated target.

References

Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2016. An actor-critic algorithm for sequence prediction. *arXiv preprint arXiv:1607.07086*.

Mauro Cettolo, Niehues Jan, Stüker Sebastian, Luisa Bentivogli, Roldano Cattoni, and Marcello Federico. 2016. The iwslt 2016 evaluation campaign. In *International Workshop on Spoken Language Translation*.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2019. The second conversational intelligence challenge (convai2). *arXiv preprint arXiv:1902.00098*.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *arXiv preprint arXiv:1905.03197*.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2019. Evaluating the state-of-the-art of end-to-end natural language generation: The E2E NLG Challenge. *arXiv preprint arXiv:1901.11528*.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109.

Tianyu He, Xu Tan, Yingce Xia, Di He, Tao Qin, Zhibo Chen, and Tie-Yan Liu. 2018. Layer-wise coordination between encoder and decoder for neural machine translation. In *Advances in Neural Information Processing Systems*, pages 7944–7954.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701.

Po-Sen Huang, Chong Wang, Sitao Huang, Dengyong Zhou, and Li Deng. 2017. Towards neural phrase-based machine translation. *arXiv preprint arXiv:1706.05565*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 605. Association for Computational Linguistics.

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing dialogue agents via meta-learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5459.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Yijun Wang, Yingce Xia, Li Zhao, Jiang Bian, Tao Qin, Guiquan Liu, and Tie-Yan Liu. 2018. Dual transfer learning for neural machine translation with

CNN / Daily Mail			
Models	***ROUGE 1***	***ROUGE 2***	***ROUGE L***
GPT Finetune	37.4	18.1	27.7
w/o Pre-Train	35.5	17	26.2
VLM mutli-task	36.6	17.7	27
VLM-10 (+ DIst.)	35.0 (36.2)	16.5 (17.3)	25.0 (25.7)
VLM-50 (+ DIst.)	36.4 (36.8)	17.5 (17.9)	26.6 (26.8)
VLM-100 (+ DIst.)	36.5 (37.0)	17.6 (18.0)	27.0 (27.0)
VLM-300 (+ DIst.)	36.6 (36.7)	17.6 (17.7)	26.6 (26.7)
PGNet (See et al., 2017)	39.53	17.28	36.38
Bottom-Up (Gehrmann et al., 2018)	41.22	18.68	38.34
UniLM (Dong et al., 2019)	43.33	20.21	40.51
T5-11B (Raffel et al., 2019)	43.52	21.55	40.69

Table 1: Summarization results.

Persona			
Models	***Perplexity***	***BLEU***	***Consistency (C)***
GPT Finetune	13.13	2.17	0.71
w/o Pre-Train	37.77	0.99	0.12
VLM mutli-task	13.15	0.84	0.27
VLM-10	15.76	1.63	0.86
VLM-50	14.54	1.84	0.72
VLM-100 (+ DIst.)	14.06 (89.34)	1.99 (2.15)	0.76 (0.72)
VLM-300	13.73	1.98	0.74
Deep Copy (Yavuz et al., 2019)	54.58	4.09	-
PAML-TRS (Madotto et al., 2019)	30.42	1.0	0.07
ADAPT Centre (ConvAI2) (Dinan et al., 2019)	29.85	-	-
Persona-Chat (Zhang et al., 2018)	35.07	-	-
TransferTransfero (Wolf et al., 2019)	17.51	-	-

Table 2: Persona Chat results.

marginal distribution regularization. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Felix Wu, Angela Fan, Alexei Baevski, Yann N Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. *arXiv preprint arXiv:1901.10430*.

Semih Yavuz, Abhinav Rastogi, Guan-Lin Chao, and Dilek Hakkani-Tur. 2019. Deepcopy: Grounded response generation with hierarchical pointer networks. *arXiv preprint arXiv:1908.10731*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

CoQA	
Models	***F1***
GPT Finetune	67.7
w/o Pre-Train	15.1
VLM mutli-task	69.3
VLM-50 (+ DIst.)	55.8 (56.2)
VLM-100 (+ DIst.)	64.3 (62.9)
VLM-300 (+ DIst.)	66.2 (64.8)
Seq2Seq (Reddy et al., 2019)	27.5
PGNet (Reddy et al., 2019)	45.4
DrQA (Reddy et al., 2019)	54.7
UNILM (Dong et al., 2019)	82.5
Human (Reddy et al., 2019)	89.8

Table 3: CoQA results.

NMT	
Models	**BLUE**
GPT Finetune	25.45
w/o Pre-Train	16.52
VLM mutli-task	22.49
VLM-10 (+ DIst.)	6.27(12.57)
VLM-50 (+ DIst.)	14.79(20.09)
VLM-100 (+ DIst.)	19.89(22.39)
VLM-300 (+ DIst.)	23.77(24.19)
Transformer (Vaswani et al., 2017)	29.2
DynamicConv (Wu et al., 2019)	35
MIXER (Ranzato et al., 2015)	21.83
AC+LL (Bahdanau et al., 2016)	28.53
NPMT (Huang et al., 2017)	28.96
Dual Transfer Learning (Wang et al., 2018)	32.35
LYC Transforemer (He et al., 2018)	35.07

Table 4: NMT results.

NLG						
Models	***BLEU***	***NIST/10***	***METEOR***	***ROUGE L***	***CIDEr/10***	***norm. avg.***
GPT Finetune	66.44	0.85279	0.4548	0.6911	0.22546	57.771
w/o Pre-Train	60.54	0.81697	0.4152	0.6471	0.19086	53.5106
VLM mutli-task	65.63	0.8342	0.4525	0.6889	0.22213	57.0806
VLM-10	67.1	0.85046	0.4545	0.6935	0.229	57.9692
VLM-50	66.01	0.84124	0.4568	0.6876	0.22128	57.3404
VLM-100	65.38	0.83922	0.4564	0.6893	0.21972	57.1688
VLM-300	66.18	0.84876	0.4539	0.6897	0.22387	57.5606
VLM-10 + DIst.	65.03	0.83199	0.456	0.6849	0.21286	56.721
VLM-50 + DIst.	65.23	0.83326	0.4576	0.6866	0.21287	56.8526
VLM-100 + DIst.	64.35	0.82485	0.4584	0.6852	0.20989	56.4368
VLM-300 + DIst.	65.19	0.83481	0.4575	0.6878	0.21182	56.8766
TGEN baseline (Dušek et al., 2019)	65.93	0.86094	0.4483	0.685	0.22338	57.5384
SLUG (Dušek et al., 2019)	66.19	0.8613	0.4454	0.6772	0.22615	57.44

Table 5: NLG results.

NMT IWSLT 2014	
Source	Wenn ihr mit jemanden in den 20ern arbeitet, einen liebt, wegen einem in den 20ern Schlaf verliert, ich möchte euch seh en... O.k. Großartig. Leute in den 20ern sind wirklich wichtig.
GPT-2 Finetune	If you work with somebody in the ’20s, you love them because you lost a loved one in the ’20s, I want to see you – – great. People in the ’20s are really important.
VLM	If you work with somebody in the ’20s, because of a love lost in the ’20s, I want to see you – OK. Great. People in the ’20s are really important.
LM-Head	If you work with someone in the 20ern, you love,
Target	If you work with twentysomethings, you love a twentysomething, you’re losing sleep over twentysomethings, I want to see — Okay. Awesome, twentysomethings really matter.
Source	Ja, die Leute lassen sich später häuslich nieder als früher, aber das machte Alex’ 20er nicht zum Entwicklungsausfall.
GPT-2 Finetune	Yes, people will be more domestic in the future than they used to be, but that didn’t make Alex’ 20s for failure.
VLM	Yes, people would come up later than they used to, but that didn’t make Alex’ 20s a disaster.
LM-Head	Yes, people are later going to come back as former former, but that doesn’t make Alex’ 20s anymore.
Target	Yes, people settle down later than they used to, but that didn’t make Alex’s 20s a developmental downtime.
Source	Leute in den 20ern wie Alex und ich hatten nichts als Zeit.
GPT-2 Finetune	People in the ’20s like Alex and I didn’t have time for time.
VLM	People in the ’20s like Alex and I had nothing but time.
LM-Head	People like Alex and I had nothing as a time.
Target	Twentysomethings like Alex and I had nothing but time.

Table 6: NMT examples.

Summurization CNN Daily Mail	
Source	los angeles -lrb- cnn -rrb- it’s more than just one state’s internal problem. the historic california drought hurts the,rest of the union, too. that’s because california is a breadbasket to the nation, growing more than a third of its vegetables and nearly two-thirds of its fruits and nuts. here’s why we should heed the ongoing drought in the most populous state, a slowly expanding natural disaster now in its fourth year that this week prompted gov. jerry brown to announce a,mandatory 25 % ....
GPT-2 Finetune	the drought hurts the rest of the union, too, because california is a breadbasket to the nation, the drought is, causing the nation’s price to rise, thanks to a strong dollar.
VLM	california drought hurts rest of the union, too, says jerry brown. brown: prices rose last year for these items on your kitchen table. he says prices expected to rise this year because of inflation.
LM-Head	new : the drought has hurt the rest of the union, the drought has hurt the rest of the union, too, the drought has hurt the rest of the union, too, the drought has hurt the rest of the union, too.
Target	americans paid more for some fruits and vegetables last year because of the drought, tourists will now have to ask for a glass of water at a california restaurant, perhaps the only good thing is another “ great ” wine grape harvest last year.
Source	-lrb- cnn -rrb- governments around the world are using the threat of terrorism – real or perceived – to advance executions, amnesty international alleges in its annual report on the death penalty. “ the dark trend of governments using the death penalty in a futile attempt to tackle real or imaginary threats to state security and public safety was stark last year, ” said salil shetty, amnesty’s secretary general in a release. “ it is shameful that so many states around the world are essentially playing with people’s lives – putting people to death for ‘ terrorism’or to quell internal instability on the ill-conceived premise of deterrence.
GPT-2 Finetune	report: governments are using the threat of terrorism to advance executions, china is mentioned as having used, the death penalty as a tool in its “ strike hard ” campaign., the number of executions worldwide has gone down by almost 22 % on the previous year.
VLM	amnesty international releases its annual report on the death penalty, the report cites the example of pakistan, lifting a six-year moratorium on the execution of civilians.
LM-Head	amnesty international : death penalty is a tool in state-sanctioned killing campaigns, china is also mentioned, as having used the death penalty as a tool in its “ strike hard ” campaign against terrorism.,china is also mentioned, as having used the death penalty as a tool in its “ strike hard ” campaign against terrorism
Target	amnesty’s annual death penalty report catalogs encouraging signs, but setbacks in numbers of those sentenced to death.,organization claims that governments around the world are using the threat of terrorism to advance executions., the number of executions worldwide has gone down by almost 22 % compared with 2013, but death sentences up by 28 %.

Table 7: SUM examples.

Conversational QA (CoQA)	
Source	(CNN) – Dennis Farina, the dapper, mustachioed cop turned-actor best known for his tough-as-nails work in such TV series as ”Law Order,” ”Crime Story,” and ”Miami Vice,” has died. He was 69. ”We are deeply saddened by the loss,of a great actor and a wonderful man,” said his publicist, Lori De Waal, in a statement Monday. ” Dennis Farina was always warmhearted and professional, with a great sense of humor and passion for his profession. He will be greatly missed by his family, friends and colleagues.” Farina, who had a long career as a police officer in Chicago, got into acting through director Michael Mann, who used him as a consultant and cast him in his 1981 movie,”Thief.” That role led to others in such Mann-created shows as ”Miami Vice” (in which Farina played a mobster) and ”Crime Story” (in which he starred as Lt. Mike Torello). Farina also had roles, generally as either cops or gangsters, in a number of movies, including ”Midnight Run” (1988), ”Get Shorty” (1995), ”The Mod Squad” (1999) and ”Snatch” (2000). In 2004, he joined the cast of the long-running ”Law Order” after Jerry Orbach’s departure, playing Detective Joe Fontana, a role he reprised on the spinoff ”Trial by Jury.” Fontana was known for flashy clothes and an expensive car, a distinct counterpoint to Orbach’s rumpled Lennie Briscoe. Farina was on ”Law Order” for two years, partnered with Jesse L. Martin’s Ed Green. Martin’s character became a senior detective after Farina left the show. Q1: Is someone in showbiz? A1: yes Q2: Whom?
GPT-2 Finetune	Dennis Farina
VLM	Dennis Farina
LM-Head	Michael Mann
Target	Dennis Farina

Table 8: CQA examples.

Conversational QA (CoQA)	
Source	Document: Once upon a time, in a barn near a farm house, there lived a little white kitten named Cotton. Cotton lived high up in a nice warm place above the barn where all of the farmer's horses slept. But Cotton wasn't alone in her little,home above the barn, oh no. She shared her hay bed with her mommy and 5 other sisters. All of her sisters were cute and,fluffy, like Cotton. But she was the only white one in the bunch. The rest of her sisters were all orange with beautiful white tiger stripes like Cotton's mommy. Being different made Cotton quite sad. She often wished she looked like the rest of her family. So one day, when Cotton found a can of the old farmer's orange paint, she used it to paint herself like them. When her mommy and sisters found her they started laughing. "What are you doing, Cotton?!" "I only wanted to be,more like you". Cotton's mommy rubbed her face on Cotton's and said "Oh Cotton, but your fur is so pretty and special, like you. We would never want you to be any other way". And with that, Cotton's mommy picked her up and dropped her into,a big bucket of water. When Cotton came out she was herself again. Her sisters licked her face until Cotton's fur was all all dry. "Don't ever do that again, Cotton!" they all cried. "Next time you might mess up that pretty white fur of yours and we wouldn't want that!" Then Cotton thought, "I change my mind. I like being special". Q1: Whose paint was it? A1: the farmer Q2: What did Cotton's mother and siblings do when they saw her painted orange?
GPT-2 Finetune	laughed
VLM	laughed
LM-Head	They put her in a big bucket of water
Target	laughed

Table 9: CQA examples.

Dialogue Persona-Chat	
Source	**Persona**: i read twenty books a year i'm a stunt double as my second job. i only eat kosher. i was raised in, a single parent household. **Contex**t: USR: hello what are doing today? SYS: i am good, i just got off work and tired, i have two jobs. USR: i just got done watching a horror movie
GPT-2 Finetune	i love horror movies, i'm a stunt double.
VLM	i love horror movies, i read twenty books a year.
LM-Head	i love horror movies, i love them.
Target	i rather read, i've read about 20 books this year.
Source	**Persona**: i'm a little girl. i've superpowers. i like to make time stop. i live in the cloud. **Context**: USR: hi i dye my hair 4 times a year
GPT-2 Finetune	that is cool. i like to dye my hair.
VLM	wow that's a lot of dyeing. i love to dye.
LM-Head	i love to make time stop
Target	do you live in the sky? i do. in the clouds.

Table 10: DLG examples.

Natural Language Generation NLG	
Source	name = Blue Spice eatType = coffee shop near = Crowne Plaza Hotel customer rating = 5 out of 5
GPT-2 Finetune	Blue Spice is a coffee shop near Crowne Plaza Hotel with a 5 out of 5 customer rating.
VLM	Blue Spice is a coffee shop near Crowne Plaza Hotel. It has a customer rating of 5 out of 5.
LM-Head	Blue Spice is a coffee shop near Crowne Plaza Hotel with a customer rating of 5 out of 5.
Target	- The coffee shop Blue Spice is based near Crowne Plaza Hotel and has a high customer rating of 5 out of 5. - The Blue Spice coffee shop, near Crowne Plaza Hotel, has a customer rating of 5 out of 5. - If you want a coffee shop rated 5 out of 5 pick Blue Spice. It is located near Crowne Plaza Hotel.
Source	name = Giraffe eatType = pub familyFriendly = yes food = English near = Rainbow Vegetarian Café' area = riverside
GPT-2 Finetune	Giraffe is a family friendly pub located near Rainbow Vegetarian Café. It serves English food and is located in t he riverside area
VLM	Giraffe is a family friendly pub located near Rainbow Vegetarian Café in the riverside area.
LM-Head	Giraffe is a pub near Rainbow Vegetarian Café that serves English food. It is family friendly and located in the,riverside area.
Target	- The Giraffe,,a child-friendly, English pub, is,located in riverside area, near Rainbow Vegetarian Café. - Giraffe is a pub that is children friendly and,provides English food.It is located on the riverside near Rainbow Vegetarian Café. - There is an English based pub called Giraffe. It is located in the riverside area near the Rainbow Vegetarian Café and, yes, it is kid friendly.

Table 11: NLG examples.