CSL: A Large-scale Chinese Scientific Literature Dataset for Cross-task Evaluation

Anonymous ACL submission

Abstract

Scientific literature serves as a high-quality corpus, which could provide natural annotated data for many natural language processing (NLP) research. In this work, we introduce 005 a Chinese Scientific Literature dataset – CSL, which contains the titles, abstracts, keywords and academic fields of 400,000 papers. The rich semantic information in these scientific literature creates extensive NLP t asks and provides a natural cross-task scenario. Based on this, we present a cross-task few-shot bench-011 012 mark. To evaluate the cross-task transferability of the model, we design scenarios with 014 different aspects and difficulties. Compared with previous cross-task benchmarks, these 016 tasks are constructed from homogeneous corpus, allowing researchers to investigate the re-017 lationships between tasks, without being dis-019 turbed by heterogeneous data sources, annota-020 tion, and other factors. We analyze the behavior of existing text-to-text models on the pro-021 posed benchmark, and reveal the challenges for cross-task generalization, which provides a valuable reference for future research. Code and data are publicly available at GitHub¹.

1 Introduction

027

035

040

As the embodiment of human research knowledge, scientific literature is known as a rich source of informative data, supporting various NLP research (Luan et al., 2018; Cohan et al., 2019). So far, several scientific-related resources e.g. large-scale literature corpus (Lo et al., 2020; Saier and Färber, 2020), citation graphs (Sinha et al., 2015; Tang et al., 2008; Zhang et al., 2019), scientific downstream tasks (Lee et al., 2020; Beltagy et al., 2019) are available. Previous works, however, have primarily relied on digital libraries, such as arXiv, PubMed, CiteSeerX and ACL Anthology, which are mostly centered around the English language and focus on specific research fields.

Ihttps://github.com/CSL-Dataset/CSL_ Dataset To fill the gap of non-English scientific corpora, in this paper we introduce CSL: a large-scale Chinese Scientific Literature dataset. CSL is obtained from 1982 Chinese core journals and contains meta-information of 400,000 papers with a wide range of distribution and fine-grained discipline annotation (67 categories).

041

042

043

044

045

047

049

052

054

057

059

060

061

062

063

064

065

066

067

068

069

070

071

073

074

075

077

078

079

Scientific literature metadata contains massive corpus information, making it a natural annotated data source with the potential to provide many highquality NLP tasks. For example, predicting the title with abstract constitutes a summarization task, and predicting the discipline is a classification task. There are hundreds of such combinations. These tasks are constructed with homogeneous data, encouraging models to share knowledge across tasks.

Cross-task generalization, i.e., how to learn a new task efficiently based on the experiences of previous tasks, is an hot area in NLP community (Ye et al., 2021; Bragg et al., 2021; Sanh et al., 2021; Zhong et al., 2021). Previous studies mostly rely on heterogeneous data to create cross-task scenarios. For example, Ye et al. (2021) use 160 diverse NLP datasets to build a few-shot NLP gym; Bragg et al. (2021) use 20 dataset to construct transfer scenarios. For those cross-task scenarios above, there are multiple task-agnostic variables, such as data sources, annotation and task formats, making it difficult to reveal the relationship between specific tasks. In this paper, we introduce our cross-task benchmark, which includes a series of tasks where underlying knowledge and distribution are shared. We aim to reduce the variance of heterogeneous data and focus on evaluating connections among tasks, as well as providing a testbed for cross-task research.

We present three cross-task scenarios that are common in real applications. Each is made up of meta training tasks (meta tasks) and disjoint fewshot evaluation tasks (few-shot tasks). These scenarios show different relationships between tasks,

Title	Abstract	Keywords	Discipline	Category
城市道路绿化景观研究	随着我国城市化步伐的加快	道路; 景观; 绿化;	园艺学	农学
Research on Urban Road Greening Landscape	With the progress of urban- ization in China	Road; Landscape; Greening;	Horticulture	Agriculture
分布式库存管理 Distributed Inventory Management Strategy	分析了分布式库存的管理模型 This paper analyzes the management model of distributed inventory	分布式库存; 协调中心; Distributed inventory; Coordination center;	应用经济学 Applied Economics	经济学 Economics

Table 1: Examples of the CSL dataset.

and vary in difficulty. It allows us to better understand how the connection of tasks affects the cross-task performance of the model. We provide a prompt-based text-to-text method as our baseline, which allows full parameter sharing across different task formats and easier transfer learning. Experiment results show that text-to-text language models are capable of cross-task transfer. However, in some challenging scenarios, there is still room for improvement.

083

086

087

880

100

101

102

104

105

106

107

108

109

110

111

112

113

114

115

116

117

118

The main contributions of this paper are summarized as follows:

- We release the first large-scale Chinese Scientific Literature dataset (CSL), which can be used for many different purposes, e.g. pretraining corpus and scientific-related tasks.
- Based on the CSL, we introduce a benchmark including different scenarios for cross-task few-shot evaluation.
- We propose a prompt-based method as our baseline and the experiment results highlight the model's difficulties in learning across tasks.

2 The CSL Dataset

2.1 Data Collection and Processing

We collect Chinese papers' homepage from publicly available search engines (Wanfang Data² and CNKI³) dated 2010 to 2020, then use the XSS parser to extract the meta-information in each web page, such as the title, abstract, keyword, and the journal in which the paper was published. To improve data quality, we omit papers not published in the Catalogue of Chinese Core Periodicals, a journal evaluation standard issued by Peking University. Then, we filter out papers from comprehensive journals and only preserve papers from professional journals that focus on a specific academic field.

Category	d	len(t)	len(a)	num(k)	#
Engineering	27	19.1	210.9	4.4	177k
Science	9	20.7	254.4	4.3	35k
Agriculture	7	17.1	177.1	7.1	39k
Medicine	5	20.7	269.5	4.7	37k
Management	4	18.7	157.7	6.2	23k
Jurisprudence	4	18.9	174.4	6.1	21k
Pedagogy	3	17.7	179.4	4.3	16k
Economics	2	19.5	177.2	4.5	11k
Literature	2	18.8	158.2	8.3	10k
Art	1	17.8	170.8	5.4	5k
History	1	17.6	181.0	6.0	6k
Strategics	1	17.5	169.3	4.0	3k
Philosophy	1	18.0	176.5	8.0	7k
All	67				400k

Table 2: The statistics of CSL dataset.

According to the Catalogue, we divide academic fields into 13 categories (Engineering, Medicine, etc.) and 67 disciplines (Mechanical Engineering, Oral Medicine, etc.). Each journal is associated with a cateory and discipline, therefore papers are annotated with two classification labels according to its published journal. For example, papers from "Chinese Journal of Computers" are categorized into Engineering category and Computer Science discipline.

Finally, we collect 400K instances for CSL dataset, represented as a tuple $\langle t, a, k, c, d \rangle$, where t is the title, a is the abstract, k is a list of keywords, c is the textual category label and d is the textual discipline label. CSL covers a wide range of academic fields, Table 1 shows the concrete examples in CSL dataset and the detailed statistics are provided in Table 2.

2.2 Task Formats

The current version of the CSL dataset contains 5 columns, which constitutes $\sum_{i=1}^{5} (\sum_{j=1}^{5-i} C_5^j \times C_{5-j}^i) = 180$ different tasks, where *i* and *j* are the number of input and output fields.

Previous works (Raffel et al., 2020; Gao et al., 2020) try to unify different tasks into general format for easier transfer learning. We extend this idea by designing task-specific prompts which cast

143

144

145

²https://www.wanfangdata.com.cn

³https://www.cnki.net

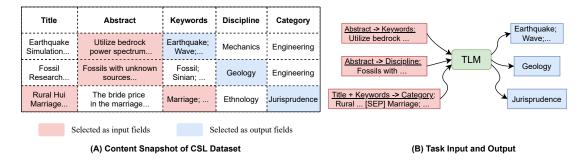


Figure 1: Overview of CSL task specific-prompts. (A) A content snapshot of the tabular CSL dataset. (B) TLM is a text-to-text language model. Prompts are added in front of input text such that different tasks can share loss function and output layers.

all CSL tasks into "text-to-text" format. For the CSL prompt, the input and output relationships are indicated by arrow characters, and multiple fields are coupled by plus signs. This prompt allows the model to predict multiple targets (i.e. one/manyto-many tasks) in a unified manner. Fig 1 gives a schematic overview of CSL prompts and tasks.

3 Cross-task Benchmark

146

147

148

149

150

151

152

153

154

155

156

157

158

159

160

161

162

163

165

166

167

168

170

171

172

173

174

175

176

Based on scientific NLP tasks derived from CSL dataset, we construct different cross-task scenarios, each of which contains pairs of meta tasks T_{meta} and disjoint few-shot test tasks T_{few} .

The model initially only has access to T_{meta} for meta training purpose, which captures the pattern in which the task structure differs from target tasks. The model's task-transferability is then evaluated by investigating the meta training stage's relative performance gain on learning T_{few} .

To comprehensively analyze the cross-task generalization, we manually design three scenarios concentrated on different aspects and varied in difficulties (i.e. the difficulty for few-shot tasks to leverage from meta tasks), each of which contains several partitions of T_{meta} and T_{few} . Tasks and their prompts are shown in Table 3. We sample 100 instances for each meta tasks and k-shot (k samples per class for classification tasks) for few-shot tasks with 8 different random seeds. 0, 1, 2, 4, 8 shot(s) are used for zero/few-shot training and 64 for validation/test.

3.1 Single-leap Bridging

177In Scenario 1, we evaluate the implicit bridging178proposed by Johnson et al. (2017) as a zero-shot179translation solution. It represents a real-world ap-180plication where there is rare training data between181the source field and the target field, an interme-

Part.	Meta Tasks	Few-shot Tasks
Scenario 1		
1-1	Abst.→Kw., Kw→Title.	Abst.→Title.
1-2	Kw.→Title, Title→Dcp.	Kw.→Dcp.
1-3	Abst. \rightarrow Title+Ctg., Title \rightarrow Kw.	Abst.+Dcp.→Kw.
Scenario 2		
2-1	Abst.→Kw., Kw.→Title Title→Ctg.	Abst. \rightarrow Ctg.
2-2	Abst.→Title., Title→Kw. Kw.→Dcp.	Abst. \rightarrow Dcp.
Scenario 3		
3-1	Kw.→Abst., Title→Dcp.	Kw.→Dcp.
3-2	Abst.→Kw., Title→Kw. Title→Ctg.	Abst. \rightarrow Ctg.

Table 3:Task prompts for each partition, Ctg: Category, Dcp: Discipline, Kw: Keywords, Abst: Abstract.

182

183

184

185

187

189

190

191

192

193

194

195

196

197

198

199

201

202

diate field can be utilized as a bridge. For example Abstract \rightarrow Keywords and Keywords \rightarrow Discipline enable a zero-shot task Abstract \rightarrow Discipline. Based on this, we design partition₁₋₁ and partition₁₋₂ representing naive single-leap implicit bridging tasks, partition₁₋₃ explores semiconnected bridging with one-to-many and many-to-one tasks.

3.2 Two-leap Bridging

Expanding the single-leap bridging, in *Scenario* 2, there are two intermediate fields between the source field and the target field, implying that three meta tasks are involved in the bridging.

Additional meta tasks bring more samples for meta training, which theoretically have the same potential upper bound as *Scenario 1*. However, achieving it requires stronger transferability, which poses a tougher challenge for few-shot learners. In *partition*₂₋₁ and *partition*₂₋₂, we design meta tasks spanning through two leaps.

3.3 Broken Bridge

In real-world scenarios, it can be difficult to locate the intermediate field that connects sources and

					Scenario	1					Scena	rio 2				Scen	ario 3	
#	Model	Т	S ₁₋₁	Cl	LS ₁₋₂	K	G1-3	Avg.		CLS ₂₋₁			CLS ₂₋₂		CL	S ₃₋₁	CL	S ₃₋₂
		few	+meta	few	+meta	few	+meta	Δ_m	few	+meta	Δ_m	few	+meta	Δ_m	+meta	Δ_m	+meta	Δ_m
1-shot	T5 _{base}	9.3	28.8	1.5	3.9	1.1	14.4	+11.7	7.9	17.0	+18.8	1.8	5.3	+3.5	4.2	+2.7	15.2	+7.3
	BART _{base}	21.9	29.9	6.3	16.9	2.8	26.1	+14.0	31.1	40.0	+8.9	11.9	10.7	-1.2	16.4	+10.1	29.1	-2.0
2-shot	T5 _{base}	11.7	28.1	1.6	4.5	2.0	16.7	+11.4	11.4	26.7	+5.6	2.6	3.3	+0.7	3.7	+2.2	23.7	+12.2
	BART _{base}	21.9	30.1	7.0	16.9	3.2	28.6	+14.5	31.2	43.3	+12.1	10.1	9.8	-0.4	16.2	+9.2	34.5	+3.3
8-shot	T5 _{base}	17.6	30.5	1.6	7.1	9.7	13.8	+7.5	38.1	40.3	+2.2	6.8	12.7	+5.9	6.7	+5.1	30.3	-7.8
	BART _{base}	26.6	31.3	17.4	20.9	22.6	30.8	+5.4	49.7	47.7	-2.0	22.3	19.4	-2.9	20.0	+2.6	30.1	-4.3

Table 4: Performance of text-to-text models on cross-task evaluation. The columns **few** presents directly finetuning on T_{few} , the **+meta** is first fine-tuning on T_{meta} (meta-tuning) and then on T_{few} . Δ_m is the Average Relative Gain of meta-tuning. We report the average metrics for each task: TS: text summarization, KG: keyword generation, CLS: classification.

targets. In *Scenario 3*, we cut off the bridges in meta tasks to imitate this condition, making the few-shot tasks even more challenging. It aims to study the impact of non-bridging homologous meta tasks on few-shot target tasks. Based on *partition*₁₋₂ and *partition*₂₋₁, we create *partition*₃₋₁ and *partition*₃₋₂ by modifying one of the meta tasks, which causes the bridge disconnected while the others unchanged.

4 Experiments

206

207

210

211

212

213

214

215

216

217

218

219

222

223

224

226

229

235

237

238

239

4.1 Baseline and Metrics

We consider different pre-trained text-to-text models including T5 (Raffel et al., 2020) and BART (Lewis et al., 2019) as our baselines. However, since there are few publically available Chinese versions of them, we conduct pre-training from scratch. Pre-training details are shown in Appendix A.

Following (Ye et al., 2021), we use multi-task fine-tuning to evaluate the above pre-trained models with and without meta tasks (i.e. meta-tuning) separately in each partition. Fine-tuning hyperparameters and other details are shown in Appendix B. We report the average of results over repeated experiments. The evaluation toolkit can be found in Appendix C.

For evaluation metrics, we adopt ROUGE-L and BLEU for summarization tasks; F1 and Bpref (Buckley and Voorhees, 2004) for keyword generation tasks. Classification tasks adopt accuracy and F1 macro as metrics. All the metrics are calculated at Chinese character level.

4.2 Results and Analysis

We observe from Table 4 that, the gain of metatuning (Δ_m) is positive on average, meaning that meta tasks generally improve task generalization.
In Scenario 1, the meta tasks boost the performance of the few-shot tasks dramatically, and this benefit is sustainable as the shots increases.

241

242

243

244

245

246

247

248

249

250

252

253

254

255

256

257

258

259

260

261

262

263

264

265

266

267

268

269

270

271

272

273

274

275

276

For Scenario 2, the benefit of meta-tuning to few-shot tasks is first noticeable, but it quickly fades with the increment of training samples and eventually drops. Finally, the models fine-tuned directly on T_{few} sometimes outperform the metatuned models. For example, *partition*₂₋₂ has the same target as *partition*₁₋₂, i.e. discipline classification, but it receives more informative input. As a result, it outperforms *partition*₁₋₂ in direct few-shot fine-tuning. However, compared with the former, it gains less from meta-tuning for the average Δ_m drops by %4.9. This demonstrates that two-leap bridging indeed increases the difficulty.

In Scenario 3, we found the Δ_m of partition₃₋₁ and partition₃₋₂ show the similar changing trend as partition₁₋₂ and partition₂₋₁. However, in comparison, they have decreased by 2.0% and 5.1% on average. The results suggest that meta-tuning on homologous tasks generally improves few-shot learning, and implicit bridging is a key factor affecting the task generalization.

More detailed experiment results and our other findings are demonstrated in Appendix D.

5 Conclusion

In this paper, we provide a large-scale Chinese Scientific Literature dataset (CSL) and use it to evaluate few-shot cross-task generalization. This represents the challenge of addressing low-resource tasks with high-resource tasks. From the experiment results, we observe that homogeneous meta tasks generally improve few-shot learners, however, when the bridge is broken, this benefit becomes negligible. We also release an open-source toolkit for extensive evaluation.

References

277

278

279

281

284

285

287

290

291

292

295

296

297

299

301

302

307 308

310

311

312

313

314

315

316

317

318

319

321

323

324

325 326

327

331

- Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3615– 3620.
- Jonathan Bragg, Arman Cohan, Kyle Lo, and Iz Beltagy. 2021. Flex: Unifying evaluation for few-shot nlp. *arXiv preprint arXiv:2107.07170*.
- Chris Buckley and Ellen M Voorhees. 2004. Retrieval evaluation with incomplete information. In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pages 25–32.
 - Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3586– 3596.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186.
- Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.
- Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019.
 Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461.
- Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S Weld. 2020. S2orc: The semantic scholar open research corpus. In *Proceedings of the*

58th Annual Meeting of the Association for Computational Linguistics, pages 4969–4983. 333

334

335

336

337

338

340

341

342

343

344

345

346

347

349

350

351

352

353

354

355

356

357

358

359

360

361

362

363

364

365

366

367

368

369

370

371

372

373

374

375

376

377

378

379

380

381

382

383

386

387

- Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.
- Tarek Saier and Michael Färber. 2020. unarxive: a large scholarly data set with publications' full-text, annotated in-text citations, and links to metadata. *Scientometrics*, 125(3):3085–3108.
- Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.
- Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, pages 243–246.
- Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. Arnetminer: extraction and mining of academic social networks. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 990–998.
- Liang Xu, Xuanwei Zhang, and Qianqian Dong. 2020. Cluecorpus2020: A large-scale chinese corpus for pre-training language model. *arXiv preprint arXiv:2003.01355*.
- Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021. Crossfit: A few-shot learning challenge for cross-task generalization in nlp. *arXiv preprint arXiv:2104.08835*.
- Fanjin Zhang, Xiao Liu, Jie Tang, Yuxiao Dong, Peiran Yao, Jie Zhang, Xiaotao Gu, Yan Wang, Bin Shao, Rui Li, et al. 2019. Oag: Toward linking large-scale heterogeneous entity graphs. In *Proceedings of the* 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 2585– 2595.
- Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

- 389 Zhe Zhao, Hui Chen, Jinbin Zhang, Wayne Xin Zhao, 390 Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. 2019. Uer: An open-source 391 toolkit for pre-training models. In Proceedings of 392 the 2019 Conference on Empirical Methods in Nat-393 394 ural Language Processing and the 9th International 395 Joint Conference on Natural Language Processing 396 (EMNLP-IJCNLP): System Demonstrations, pages 241-246. 397
- Ruiqi Zhong, Kristy Lee, Zheng Zhang, and Dan Klein.
 2021. Adapting language models for zero-shot
 learning by meta-tuning on dataset and prompt collections. In *Findings of the Association for Com- putational Linguistics: EMNLP 2021*, pages 2856–
 2878.

484

485

486

450

A Pre-training Chinese Text-to-text Language Models

404

405

406

407

408

409

410

411

412

413

414

415

416

417

418

419

420

421

422

423

424

425

426

427

428

429

430

431

432

433

434

435

436

437

438

439

440

441 442

443

444

445

446

447

448

449

For pre-training Chinese text-to-text models, we follow the architecture, optimization, and hyperparameter choices described in (Raffel et al., 2020; Lewis et al., 2019; Zhang et al., 2020). Following Chinese BERT model (Devlin et al., 2019), we use the tokenizer with a vocabulary of 21,128 Chinese characters. Models are trained for 1,000,000 steps with the sequence length of 128 and then trained for 250,000 additional steps with 512 sequence length on CLUE Corpus (Xu et al., 2020) with the batch size of 512.

B Fine-tuning Hyperparameters

For comprehensive evaluation, we additional evaluate different sizes of T5, BART and PEGASUS. When fine-tuning different pre-trained models. we use the same hyperparameters. The settings of hyperparameters are as follows. The learning rate is set to 3e-4 for T5 and 1e-5 for BART and PEGA-SUS, the batch size is 32 for training on T_{meta} and 1 for training on T_{few} . We set the number of epochs to 15 with early stopping. The maximal input and output length are set to 256, which can be shortened according to the length of the task data to speed up training. All results are reported with greedy decoding (i.e. choosing the highest-probability logit at every timestep). All experiments are conducted on 1 Tesla V100 GPU and the results are collected over average of 8 episodes. Altogether, the experiments take around 2000 GPU hours, the full results are shown in Table 9.

C Pre-training and Evaluation Toolkit

We use UER-py⁴ (Zhao et al., 2019) as our pretraining and fine-tuning platform. Based on which we implement a toolkit for cross-task evaluation including function modules: (1) Sampling Kshot examples for meta/few-shot tasks according CSL evaluation protocol. (2) Conducting metatuning on meta tasks and fine-tuning on fewshot tasks. Code, dataset and pre-trained models are available at https://github.com/ CSL-Dataset/CSL_Dataset.

D Additional Analysis and Samples

We provide full results broken down by partitions in Table 9. In this section, we describe other findings. Larger models perform better on few-shot tasks but small models are more likely to benefit from meta-training. As with previous common sense, larger models perform better on average for various tasks. However, smaller models have considerable task transferability, sometimes better than large models. For example, the average Δ_m of BART-base is 3.7%, while of BART-small is 4.4%.

Sentence-level denoising pre-training performs better on few-shot learning. In the majority of cases, BART and PEGASUS outperform T5. We speculate that it is because T5 is pre-trained with token-level denoising tasks, whereas BART and PEGASUS are pre-trained with sentence-level denoising tasks, which makes pre-training tasks closer to downstream tasks, allowing for easier few-shot transfer.

Meta-tuning on bridging tasks enables zeroshot learning. In *O-shot* rows, we present results of direct evaluating meta-tuned models on T_{few} without few-shot training, which indicates promptbased text-to-text models have zero-shot task generalization.

Large-scale fine-tuning enables multi-task language generator. We fine-tuned a T5-base model with full CSL training set and all tasks. Output samples of summarization and keyword generation tasks are shown in Table 5 and Table 6. Conditional text generation samples are shown in Table 7. For the rules of constructing CSL prompt is extensible, which makes prompts feasible at the inference phase even it is unseen during the training. For example, " $Abstract \rightarrow Abstract$ " indicates rewriting a given input abstract. Table 8 presents samples of zero-shot text rewriting task. The samples for these different tasks above all come from the same model.

⁴https://github.com/dbiir/UER-py

Prompt: 摘要->标题 Abstract -> Title

Input Text: 综述了纳米磁性粒子和磁性脂质体的制备 方法,同时简要介绍了磁性脂质体在磁性分离、靶向药 物、热疗、组织工程和造影剂等领域的应用进展.

This paper briefly discusses magnetic nanoparticles and magnetic liposomes preparation methods, as well as the progress of magnetic liposomes in the fields of magnetic separation, targeted drugs, hyperthermia, tissue engineering, and contrast agents.

Prediction: 纳米磁性粒子和磁性脂质体的制备及应用 进展

Progress in the Preparation and Application of Magnetic Manoparticles and Magnetic Liposomes

Ground Truth: 磁性脂质体的制备及应用研究进展

Progress in the Preparation and Application of Magnetic Liposomes

Prompt: 摘要->标题

Abstract -> Title

Input Text: 通过对美国职业排球运动员进行非结构性 访谈研究美国职业排球运动员对赞助商和赞助行为的 态度... 赞助商应尊重运动员的情感和观点,从而使双方 都能获得长远利益.

Through interviews, research was conducted on the attitudes of American professional volleyball players regarding sponsors and sponsorship activities ... Sponsors should respect athletes' feelings and opinions in order for both sides to profit in the long run.

Prediction: 美国职业排球运动员对赞助商和赞助行为的态度研究

Research on American Professional Volleyball Players' Attitudes Towards Sponsors and Sponsorship Behaviors

Ground Truth: 美国排球运动员对赞助的态度分析 Analysis of American Volleyball Players' Attitudes towards Sponsorship

Table 5: Samples of summarization.

 Prompt: 摘要->关键词

 Abstract -> Keywords

 Input Text: 位置跟踪是移动机器人自主导航中的一个

 主要任务...文中给出了一个基于线性系统模型的位置

 估计方法

 Position tracking is a major task in the autonomous navigation of mobile robots ... This paper proposes a position estimation method based on the linear system model.

Prediction: 移动机器人; 位置估计; 线性系统模型; 线性 近似

Mobile robots; Position estimation; Linear system model; Linear approximation

Ground Truth: 移动机器人; 位置跟踪; 卡尔曼滤波 Mobile robots; Position tracking; Kalman filter **Prompt:** 摘要->关键词

Abstract -> Keywords

Input Text: 通过对祁连山自然保护区周边农牧民经济状况的调查发现阻碍经济发展的问题... 提出了发展生态旅游等适合本地区经济发展的模式.

Problems with economic development were discovered during an investigation of the economic conditions of farmers and herders in the Qilian Mountain Nature Reserve ... Ecotourism and other models for local economic development were proposed.

Prediction: 祁连山自然保护区; 农牧民; 经济发展模式 Qilian Mountain Nature Reserve; Peasants and herdsmen; Economic development model

Ground Truth: 祁连山自然保护区; 周边经济; 发展模式 Qilian Mountain Nature Reserve; Peripheral economy; Development model

Table 6: Samples of keywords generation.

Prompt: †	示题->摘要
Title -> Ab	ostract

Input Text: 藏东南地区生态旅游资源评价与开发建议 Evaluation and Development Suggestions on Ecotourism Resources in Southeastern Tibet

Prediction: 藏东南地区是我国重要的生态旅游资源 区,其旅游资源的开发是一个重要的课题.本文从旅游资 源的评价与开发的角度,对藏东南地区生态旅游资源开 发的现状进行了分析,并提出了相应的开发建议.

Southeast Tibet is a significant ecotourism resource area in China, and the development of its tourism resources is a key issue. This paper focuses on the current of development of ecotourism resources in Southeast Tibet from the standpoint of tourism resource evaluation and development. Also, make recommendations for the future development.

Prompt: 标题->摘要 Title -> Abstract

Input Text: 促进我国开放式基金发展的七大对策 Seven Strategies to Promote the Development of China's Open Funds

Prediction: 文章分析了我国开放式基金所面临的困 难,提出了目前我国开放式基金发展中存在的弊端,提出 了对策:要加快开放式基金发展,探索新型开放式投资工 且

This paper highlighted the problems facing China's open-end funds, identifies flaws in the current development of openend funds, and suggests solutions, including accelerating the development of open-end funds and exploring new open-end investment tools.

Table 7: Samples of conditional text generation.

Prompt: 摘要->摘要 Abstract -> Abstract

Input Text: 自蔓延高温合成(shs)技术是一种材料制备的新工艺,但该技术存在的某些局限性阻碍了它的进一步发展和工业化应用.

Self-propagating high-temperature synthesis (shs) technology is a new process for material preparation, but some limitations of this technology hinder its further development and industrial application.

Prediction: 自蔓延高温合成方法是一种将非金属超细胞材料及聚醚基金属离子转换成金属基金属的新技术,它以其独特的特点而成为一种化学研究的热点.该技术在材料制备领域的应用受到科技企业的认可和研究,但它的优越性并没有显著影响其在工业上商业化应用的进一步发展.

The self-propagating high-temperature synthesis method is a new technology that converts non-metallic supercellular materials and polyether-based metal ions into metal-based metals. With its unique characteristics, it has become a hot spot in chemical research. In the field of material preparation, it is recognized and researched by technology companies, but its superiority has not significantly affected the further development of its commercial applications in industry.

Prompt: 摘要->摘要 Abstract -> Abstract

Input Text: 烟大轮渡铁路栈桥采用两跨式全焊变截面 下承式钢板梁,跨度为52.645m+29.855m.从结构形式、 提升方式等方面介绍该栈桥的十大关键性技术.

The Yantai Ferry Railway Trestle uses a two-span fully welded variable cross-section through steel plate girder with a span of 52.645m+29.855m. Ten key technologies of the trestle are introduced from the aspects of structure, lifting methods, etc.

Prediction: 烟大轮渡桥是烟西市第一座高龄全焊变截面下承式钢板梁斜拉桥.在整体展示桥梁结构、桥梁板结构等方面介绍了该桥的关键技术,进而为该桥长期保持桥梁的强度构造奠定了有利实践基础.

Yanxi City's first all-welded variable cross-section steel plate girder cable-stayed bridge is the Yantai Ferry Bridge. This paper introduces the bridge's key technologies in the context of the overall display of the bridge structure and the bridge slab structure, and then lays a favorable practical foundation for the bridge's long-term strength.

Table 8: Samples of rewriting.

									Scenario 1	1										Scenario 2	io 2						ŝ	Scenario 3		
Models Few -mem few -mem few -mem -mem 13.3 15.4 Jer Jer <td< th=""><th>Modely R. L. B. H. L. B. Eve. Tention Forse Forse Forse</th><th></th><th></th><th></th><th>TS_{1}</th><th>-</th><th></th><th></th><th>CLS₁₋₂</th><th></th><th></th><th></th><th>KG₁₋₃</th><th></th><th></th><th></th><th></th><th>CLS_{2-1}</th><th></th><th></th><th></th><th>0</th><th>LS_{2-2}</th><th></th><th></th><th>CLS</th><th>S₃₋₂</th><th></th><th>CLS₃</th><th>5</th></td<>	Modely R. L. B. H. L. B. Eve. Tention Forse Forse Forse				TS_{1}	-			CLS ₁₋₂				KG ₁₋₃					CLS_{2-1}				0	LS_{2-2}			CLS	S ₃₋₂		CLS ₃	5
Marrow R.L. B4 R.L. B4 Acc FI Acc	$ \begin{array}{cccccccccccccccccccccccccccccccccccc$		Modele	fev	~	+me	ta	few		+meta		few	+	meta	Avg.		ew	+mć	sta		few		+meta			+meta		Ŧ	meta	
T5 and T5 and			SIDDOTAT	R-L	B-4	R-L	B-4							Bpr.			F1	Acc	F1	Δ_m	Acc									Δ_m
Thim Thim Thi To <	Martinal 141 17 129 114 7 129 114 7 129 114 7 129 139 333 305 113 313 334 333 334 333 333 334 333 334 333 334 333 334 333 334 334 333 334 334 334 334 334 334 334 334 334 334 334 334 334 333 334 347 334 347 333 344 333 334 347 333 344 333 344 333 344 333 347 347 343 343 343 343 343 343		$T5_{small}$			37.9	12.4		3.		8		1.4	3.8				10.3	6.1			4		.1						
	MARTime 394 153 103 133 300 111 139 133 301 133 133 301 133 133 303 133 133 333 301 133 133 333 333 334 133 333 334 133 333 334 133 333 334 133 334 343 344 343 344 34		$T5_{small}$			14.1	1.7		1		1.4		0.8	0.8				16.7	15.3			. 1		œ.						
	Method 365 11 10 23 11 23 20 13 23 20 13 14 14 13 <	0-shot	$BART_{base}$			39.4	15.3		11		8.3		2.6	3.1				33.3	30.0			-		3.8						
PEGA _{less} 370 109 239 213 16 311 371 393 357 393 357 353 351 353 351 353 351 353 351 353 351 353 351 353 351 353 351 353 351 353 351 353 351 353 353 353 351 353	$ \begin{array}{ c c c c c c c c c c c c c c c c c c c$		BART _{small}			36.5	11.1		15		9.1		2.3	4.2				34.6	25.4			-		1.6						
Avenge 330 103 156 146 117 300 133 54 215 56 32 411 21 33 51 13 71 55 56 35 45 38 27 38 27 55 53 45 38 27 45 38 27 45 38 27 45 38 27 45 38 27 45 38 27 45 45 45 45 45 46 47 45 45 46 47 45 45 46 47 45 45 46 47 45 46 47 45 45 46 47 46 47 45 47 46 47	Worklie 330 103 156 16 17 30 254 21.5 56 25.5 45.5 <th></th> <th>PEGA.base</th> <th></th> <th></th> <th>37.0</th> <th>10.9</th> <th></th> <th>2</th> <th></th> <th>1.3</th> <th></th> <th>1.6</th> <th>3.1</th> <th></th> <th></th> <th></th> <th>37.2</th> <th>30.9</th> <th></th> <th></th> <th>-</th> <th></th> <th>5.4</th> <th></th> <th></th> <th></th> <th></th> <th></th> <th></th>		PEGA.base			37.0	10.9		2		1.3		1.6	3.1				37.2	30.9			-		5.4						
			Average			33.0	10.3		1		4.6		1.7	3.0				26.4	21.5					S						
TSmull 58 0.3 347 115 28 23 126 143 101 23 47 43 43 441 43 43 441 43 43 441 43 43 441 43 43 441 43 441 43 441 441 453 43 43 441 441 441 443 453 443 54 448 441 104 153 501 443 553 301 153 441 441 533 301 553 443 54 453 301 153 441 441 533 401 401 153 401 403 401<	Themal 55 51 415 23 347 115 28 23 126 47 44 455 47 456 47 455 457 410 133 233 236 143 133 233 236 143 133 237 131 277 101 233 230 125 237 211 237 201 239 231 437 441 643 473 444 443 233 201 533 401 401 555 301 153 533 100 333 233 411 413 433 43 413 41 401 553 301 543 44 453 133 701 403 403 413 41 420 433 414 420 533 301 543 403 413 403 413 414 423 533 404 413 533 501 501 501		T5 _{base}	16.9	1.6	40.4	17.1							22.9			6.2	19.9		+9.1	2.3								13.1	+7.3
$ \begin{array}{ c c c c c c c c c c c c c c c c c c c$	AMRTers 11 127 odd 93 323 343 414 413 413 414 413 414 413 414 </td <th></th> <td>$T5_{small}$</td> <td>5.8</td> <td>0.3</td> <td>34.7</td> <td>11.5</td> <td>2.8 2</td> <td>1.3 L</td> <td></td> <td>1.8 0</td> <td>1 0.0</td> <td></td> <td>15.1</td> <td>+13.0</td> <td></td> <td>9.5</td> <td>27.2</td> <td></td> <td>+15.2</td> <td>2.4</td> <td>1.3 4</td> <td>1.7 4</td> <td>.3 +2</td> <td>.6 4.</td> <td>7 4.5</td> <td></td> <td></td> <td></td> <td>+3.3</td>		$T5_{small}$	5.8	0.3	34.7	11.5	2.8 2	1.3 L		1.8 0	1 0.0		15.1	+13.0		9.5	27.2		+15.2	2.4	1.3 4	1.7 4	.3 +2	.6 4.	7 4.5				+3.3
	ART Mail 11 12 33 17 71 340 33 471 35 417 774 494 33 471 473 474 474 474 474 474 474 474 474 474 475 475 475 375 416 513 554 416 513 554 416 513 554 475 376 307 319 34 41 34 473 473 306 393 105 32 301 155 303 305 51 474 473 305 305 34 473 473 306 473 473 301 153 301 153 301 153 304 173 140 302 154 133 154 135 155 135 154 135 135 135 136 137 140 133 131 132 134 133 135 135 <th< th=""><th>1-shot</th><th>$BART_{base}$</th><th>31.1</th><th>12.7</th><th>40.1</th><th>19.7</th><th></th><th></th><th></th><th></th><th></th><th></th><th></th><th></th><th></th><th>28.9</th><th>44.2</th><th>·</th><th>+8.9</th><th>13.7</th><th></th><th>2.3 9</th><th>.0 -1</th><th></th><th></th><th></th><th></th><th>24.9</th><th>-2.0</th></th<>	1-shot	$BART_{base}$	31.1	12.7	40.1	19.7										28.9	44.2	·	+8.9	13.7		2.3 9	.0 -1					24.9	-2.0
	EGA _{hus} 1121142621818214425922460137253447455416513453448207166159164415233201554478Verenge25206315109307165233071652330716523306719910671317315191964141731519196414128101553306Knunn310129202001317710631731341342444347430414473514473314474474Knunn31012940220701055811523358144264733041447476115474474474474474ARTNas310129402401005581152335814433432433432433432447474ARTNas31012940356641010000277221385319410407519414474ARTNas31312344334343333432443436447474ARTNas313155405201413366610277221		BART _{small}	31.0	12.3	38.3	17.5										31.3	47.1		+8.6	14.1									+2.7
Average 522 96 992 175 81 59 160 137 25 71 99 770 91 79 115 134 116 $+55$ TSmass 210 23 397 165 23 10 106 11 21 73 292 24 19 96 +14 134 116 +55 TSmass 210 23 319 105 410 38 153 140 34 55 14 153 41 25 115 353 44 155 115 353 415 134 138 153 141 158 353 413 354 49 173 159 135 116 55 133 150 135 141 128 159 135 141 138 149 116 159 149 149 149 149 149 149 1416 156 133 134	Weenage 522 96 372 13 315 116 15,5 306 13,4 116 +5,5 306 13,4 116 +5,5 306 13,4 116 +5,5 306 13,4 116 +5,5 306 13,4 116 +5,5 306 13,4 +12,2 260 13,4 +12,2 260 15,4 +12,3 15,8 41,4 12,8 13,3 13,1 13,9 13,3 14,3 13,3 13,3 14,3 13,3 14,4 13,3 14,4 14,3 13,3 14,4 14,4		PEGA.base	41.1	21.1	42.6	21.8									45.5	41.6	51.3		+4.8	20.7									+0.2
$ T_{\text{Newer}} \ \ \ \ \ \ \ \ \ \ \ \ \$			Average	25.2	9.6	39.2	17.5							27.0			23.5	37.9		+9.3	<u>`</u>									+2.3
$ TS_{mull} \begin{tabular}{lllllllllllllllllllllllllllllllllll$			T5 _{base}	21.0	2.3	39.7	16.5		.9 5.	1.				27.1	+11.4		10.1	29.2		+15.2	3.4									+12.2
$ \begin{array}{ c c c c c c c c c c c c c c c c c c c$	AMRT _{main} 310 129 402 200 87 53 184 155 11 54 155 1445 355 417 1445 358 266 474 392 110 128 84 112 83 161 153 153 153 151 153 333 411 333 413 233 413 233 413 233 413 233 413 233 413 233 24 493 133 24 474 473 513 143 335 51 433 353 113 265 344 473 50 343 353 113 213 335 511 738 513 739 518 53 56 64 910 501 64 64 64 66 60 416 304 733 531 739 511 738 749 739 131 739 147 739 147		$T5_{small}$	9.1	0.6	31.5	10.9							17.0			7.7	26.9		+16.3										+4.9
	AKT mult 318 12.3 40.2 19.7 100 6.5 18.1 15.2 3.8 4.8 14.2 30.8 2.3.3 41.3 32.4 4.9.8 13.0 9.4 12.4 7.6 -1.2 19.1 15.9 +9.2 41.0 FEGAbase 42.6 20.2 42.3 22.1 165 12.5 7.4 13.4 53.8 4.68 4.81 42.6 47.8 39.2 1.9 19.8 15.5 10.4 21.9 9.10 10.5 13.9 10.2 30.4 12.4 4.83 35.7 13.8 4.10 30.5 4.10 30.5 4.10 30.4 13.7 13.9 13.0 4.11 9.0 3.1 5.5 4.5 3.0 4.43 3.8 15.7 4.3 9.13 13.3 13.8 4.47 4.89 5.1 4.55 4.55 13.8 4.47 4.89 5.1 4.85 5.1 8.5 8.1 4.73 8.1	2-shot	$BART_{base}$	31.0	12.9	40.2	20.0	8.7 5									26.6	47.4		+12.1										+3.3
PEGA-base 426 202 423 221 165 173 224 134 338 468 48.1 426 478 392 19.8 15.2 18.5 15.6 -0.4 24.3 21.4 48.4 Average 27.1 97 38.8 17.8 8.1 5.4 15.4 13.1 2.6 7.9 13.3 410.3 10.4 7.5 10.8 8.1 -0.5 13.9 12.0 46.2 TSume 255 3.4 4.32 19.8 5.1 4.0 3.0 4.75 10.8 8.1 -0.5 13.9 12.0 40.4 TSume 31.3 12.7 9.1 7.4 4.8 4.7 8.0 5.1 4.8 4.8 4.7 8.9 3.10 4.10 3.7 2.1 7.8 2.1 18.4 4.7 8.9 5.1 4.8 7.8 4.9 5.1 5.1 5.1 5.1 5.1 5.1 5.1 <td>EEdAtase 2/2 2/2 2/2 1/2 2/2 1/</td> <th></th> <td>BART_{small}</td> <td>31.8</td> <td>12.3</td> <td>40.2</td> <td>19.7</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td>23.3</td> <td>41.3</td> <td></td> <td>+9.8</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td>+10.1</td>	EEdAtase 2/2 2/2 2/2 1/2 2/2 1/		BART _{small}	31.8	12.3	40.2	19.7										23.3	41.3		+9.8										+10.1
Average 27.1 9.7 38.8 17.8 8.1 5.4 13.1 2.6 7.9 10.2 30.6 +10.9 35.0 +10.3 10.4 7.5 10.8 8.1 -0.5 13.9 12.0 6.2 T5huse 255 3.4 43.2 19.8 5.2 3.9 5.1 4.3 0.3 5.6 4.6 21.9 9.0 32.1 27.0 41.0 35.0 8.1 4.8 38 15.7 13.8 10.4 3.9 2.8 -1.2 T5huse 7.3 0.3 35.5 11.7 5.4 3.9 13.7 13.8 4.0 13.6 4.0 3.9 5.1 4.8 4.7 8.9 5.1 4.8 4.7 8.9 5.1 4.8 5.1 18.8 4.7 8.9 5.1 4.9 5.1 8.8 5.1 18.8 4.7 8.9 5.1 5.8 5.1 5.8 5.1 5.8 5.1 5.8	Wenge 27.1 9.7 3.8 1.8 8.1 5.4 1.3.1 2.6 7.9 10.2 3.6 +10.3 10.4 7.5 10.8 8.1 +0.5 13.9 12.0 +6.2 33.7 STaue 265 3.4 43.2 198 5.2 3.9 5.1 4.3 0.3 5.6 4.6 21.9 4.00 32.1 2.7 10.3 8.1 +10.4 13.9 12.0 +6.2 33.4 555 3.4 3.5 11.7 5.2 3.9 5.1 4.8 7.7 9.1 3.0 +1.7 8.0 5.1 3.9 5.1 1.3 2.4 3.0 1.12 2.1 1.8 8.1 1.3 3.3 3.1 2.1 7.9 3.1 2.1 3.2 2.1 1.8 4.7 480 5.1 4.66 5.1 3.0 5.1 3.0 4.1 5.0 3.0 4.1 5.1 3.0 5.1		PEGA.base	42.6	20.2	42.3	22.1								-	48.1	42.6	47.8		-1.9	19.8									-1.7
$ T5_{\text{base}} \ \ \ \ \ \ \ \ \ \ \ \ \$	Type 265 34 432 198 5.1 4.3 0.3 5.6 4.6 21.9 +900 32.1 27.0 410 35.0 13.8 410.4 39 2.8 -1.2 42.6 Tyme 7.3 0.3 35.5 11.7 5.4 3.9 13.7 119 0.1 0.0 31 23.3 +14.0 0.7 8.5 8.7 7.9 -1.3 6.6 6.0 +16 30.4 ARTVase 31.3 13.5 10.3 5.5 13.7 +85 +2.5 25.6 21.3 2.6 6.0 +16 30.4 Shertrase 452 23.6 11.7 8.2 13.1 13.8 44.4 +89 55.1 467 56.1 49.6 49.4 56.1 46.7 56.1 49.6 57.4 21.8 2.4 23.5 21.3 32.9 21.9 14.9 46.9 55.4 55.4 55.4 21.8 54.4 55.4 21.8 54.4 55.4 21.8 54.4 29.3 21.3 21.3 <th></th> <td>Average</td> <td>27.1</td> <td>9.7</td> <td>38.8</td> <td>17.8</td> <td>8.1 5</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td>22.1</td> <td>38.5</td> <td></td> <td>+10.3</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td>27.6</td> <td>+5.8</td>		Average	27.1	9.7	38.8	17.8	8.1 5									22.1	38.5		+10.3									27.6	+5.8
T5sual 7.3 0.3 35.5 11.7 5.4 3.9 13.7 10.0 3.1 23.3 +13.7 20.5 31.0 +14.0 10.7 8.5 8.7 7.9 -1.3 6.6 6.0 +1.6 9.1 BARTyase 31.3 13.5 40.6 20.2 22.1 17.8 25.6 13.1 7.8 2.9 2.3 20.0 +1.7 BARTyase 31.3 12.7 39.5 18.5 2.8 15.1 13.8 44.4 +8.9 55.1 48.5 +5.5 23.6 13.7 20.9 23.6 13.3 37.8 40.4 55.4 49.7 61.5 54.5 45.5 28.3 21.1 17.8 23.7 23.7 12.5 33.2 13.3 7.4 49.7 48.7 56.1 49.7 61.5 54.5 54.5 24.5 21.8 24.9 64 6.0 41.7 PEGA-base 28.3 10.7 40.4 49.7	T3 0.3 355 117 5.4 39 13.7 11.9 0.1 0.0 3.1 23.3 +13.7 20.5 19.5 10.7 8.5 8.7 7.9 -1.3 6.6 6.0 +1.6 30.4 SARTwain 31.3 13.5 406 202 22.1 17.8 22.6 19.1 5.8 15.1 13.8 44.4 +8.9 55.1 48.5 5.5 25.5 25.5 21.3 22.6 18.6 -0.4 50.0 +1.7 56.1 BCGAbase 45.2 23.6 43.3 22.8 15.1 13.8 44.4 +8.9 55.1 45.7 54.5 45.7 54.5 25.0 21.3 22.4 21.8 24.4 21.8 24.4 70 40.7 56.1 49.6 54.5 45.7 54.5 25.3 25.1 13.8 400 45.9 55.4 54.5 54.5 25.4 25.4 21.8 60 45.9 55.4 55.7 55.7 55.4 55.4 55.7 55.7 55.7 55.7		T5 _{base}	26.5	3.4	43.2	19.8							21.9		32.1	27.0	41.0		+8.4								42.6		+9.8
BART _{base} 31.3 13.5 40.6 202 22.1 17.8 22.6 19.1 5.8 21.0 14.8 47.7 +8.9 51.9 46.8 55.1 48.5 +2.5 25.6 21.3 22.6 18.6 -2.9 23.1 18.6 +0.4 BART _{staull} 31.3 12.7 39.5 18.5 21.1 17.8 22.5 18.5 5.8 15.1 13.8 44.4 +8.9 55.2 46.7 56.1 49.6 +3.4 24.8 209 23.6 19.5 -1.3 22.3 20.0 +1.7 PEGA _{base} 45.2 23.6 43.3 22.8 23.9 23.6 19.5 1.3 22.3 20.0 4 2.4 2.4 25.7 23.6 43.3 22.8 23.9 23.6 19.5 1.3 22.3 20.0 4 2.4 2.4 25.4 21.8 27.9 74 42.4 18.7 2.4 0.8 8.1 6.1 3.0 16.4 4.9 22.7 7.5 40.7 35.5 43.9 36.7 +2.2 7.5 6.1 13.3 12.1 55.9 8.2 5.2 5.1 51.1 75 _{base} 27.9 7.4 42.4 18.7 2.4 0.8 8.1 6.1 3.0 16.4 4.9 22.7 7.5 40.7 35.5 43.9 36.7 +2.2 7.5 6.1 13.3 12.1 55.9 8.2 5.2 5.2 5.1 51.1 75 _{base} 14.8 12 31.8 11.8 4.0 2.4 10.9 9.2 0.2 3.5 26.1 11.8 25.3 20.8 32.4 43.7 6.8 19.0 15.9 19.9 16.9 4.9 16.1 13.8 4.0 2.4 12.8 11.8 4.0 2.4 10.9 9.2 0.2 0.2 3.5 26.1 11.8 25.3 20.8 32.4 28.5 7.7 6.1 13.3 12.1 5.9 8.2 5.2 5.1 51.1 75 _{base} 14.8 12 31.1 21.0 16.5 23.4 19.3 10.2 19.8 12.8 45.4 7.1 51.6 46.3 53.2 46.2 43.1 2.0 24.5 21.1 13.3 12.1 5.9 8.2 5.2 5.2 5.1 51.1 75.0 24.5 21.1 23.0 19.9 16.9 10.9 15.0 19.0 10.9 9.2 02.2 3.5 26.1 11.8 25.3 20.8 32.4 42.2 43.1 -2.0 24.5 21.1 13.3 12.1 6.7 0.9 15.0 15.4 42.5 8.5 7.2 19.8 11.8 4.0 2.4 10.9 9.2 02.2 0.2 3.5 26.1 11.8 25.3 20.8 32.4 42.2 43.1 -2.0 24.5 20.1 21.6 17.1 2.9 21.2 18.7 4.0 24.1 12.4 19.3 10.2 19.8 12.8 45.4 7.1 51.6 46.3 53.2 46.2 4.0 24.5 20.1 21.6 17.1 2.9 21.2 18.7 4.0 24.1 22.1 13.1 2.0 16.5 23.4 23.1 4.3 30.4 16.5 45.1 45.4 47.1 51.6 46.3 53.2 46.2 4.0 24.5 20.1 21.6 17.1 2.2 9 21.2 18.7 4.0 24.1 22.1 13.1 2.0 16.5 23.4 21.1 31.3 26.1 15.4 34.3 13.3 42.3 42.3 42.3 46.2 4.0 55.1 4.3 31.7 2.7 27.6 23.1 4.3 25.7 22.8 400 PEGA _{base} 145 23.1 43.9 13.2 40.3 19.2 14.7 11.4 19.3 15.9 8.7 20.2 10.2 36.3 46.3 45.4 47.1 51.0 41.9 2.4 17.2 19.7 16.7 0.6 16.4 13.9 12.7 22.8 400 23.1 4.3 25.7 22.8 400 23.1 4.3 25.7 22.8 400 23.1 4.3 25.7 22.8 400 23.1 4.3 25.7 22.8 400 23.1 4.3 25.7 22.8 400 23.1 4.3 25.7 22.8 400 23.1 4.3 25.7 22.8 400 23.1 4.3 25.7 22.1 18.8 4.1 22.2 21.1 10.0 15.9 10.1 1	3ART _{main} 313 13.5 40.6 20.2 22.1 17.8 22.6 51.9 46.8 55.1 48.5 +2.5 25.6 21.3 22.6 13.8 6 +0.4 50.0 BART _{main} 31.3 12.7 39.5 18.5 5.8 15.1 13.8 44.4 +8.9 55.1 45.5 5.5.1 23.6 21.3 27.9 23.7 12.5 33.2 13.3 37.8 +0.4 56.1 49.6 +3.4 24.8 20.9 23.6 1.7 56.1 BEGA _{base} 28.3 10.7 40.4 18.6 16.4 49 35.7 45.4 57.5 29.3 25.1 13.8 40.6 55.4 55.1 38.7 56.1 49.6 43.7 46.9 55.4 55.1 38.7 56.1 37.6 19.7 56.1 37.8 40.6 56.1 49.6 43.7 56.1 36.7 45.7 56.1 13.7 40.9 57.4 57.4 56.1 36.7 45.7 57.6 11.6 57.2 45.7 57.6 </td <th>4 chot</th> <td>$T5_{small}$</td> <td>7.3</td> <td>0.3</td> <td>35.5</td> <td>11.7</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td>23.3</td> <td></td> <td></td> <td>19.5</td> <td>36.9</td> <td></td> <td>+14.0</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td>+8.4</td>	4 chot	$T5_{small}$	7.3	0.3	35.5	11.7							23.3			19.5	36.9		+14.0										+8.4
BART small 31.3 12.7 39.5 18.5 21.1 17.8 22.5 18.5 5.8 15.1 13.8 44.4 +8.9 55.2 46.7 56.1 49.6 +3.4 24.8 209 23.6 19.5 -1.3 22.3 20.0 +1.7 PEGA _{base} 45.2 23.6 43.3 22.8 23.9 23.6 16.4 13.4 15.5 4.9 15.5 33.2 13.3 37.8 +0.4 55.4 49.7 61.5 54.5 5.5 29.3 25.1 28.8 24.9 0.4 25.4 21.8 -2.4 2.5 27.9 7.4 42.4 18.7 2.4 0.8 8.1 6.1 3.0 16.4 4.9 22.7 7.5 40.7 35.5 43.9 36.7 +2.2 7.5 6.1 13.3 12.1 5.9 8.2 5.2 5.2 5.5 5.5 5.5 1.4 2.1 13.8 10.0 7 35.8 11.8 4.0 2.4 10.9 9.2 0.2 0.2 0.2 3.5 26.1 +11.8 25.3 20.8 32.4 23.7 1.6 9.6 13.7 12.4 4.2.5 8.5 5.7 4.0 15.8 24.0 0.9 15.0 9.9 16.9 4.0 12.1 13.8 +0.0 7 35.8 11.8 4.0 2.4 10.9 9.2 0.2 0.2 3.5 26.1 +11.8 25.3 20.8 32.4 23.7 1.6 9.6 13.7 12.4 4.2.5 4.5 3.7 +0.9 BART _{base} 35.8 17.5 41.4 21.2 19.5 15.3 22.9 18.9 14.8 30.4 16.5 45.1 57.4 52.6 46.8 52.2 43.1 -2.0 24.5 20.1 21.6 17.1 -2.9 21.2 18.7 +0.9 BART _{small} 36.6 16.9 41.8 21.1 21.0 16.5 23.4 19.3 10.2 19.8 12.8 45.4 7.1 51.6 46.3 53.2 46.2 4.3 1.2 2.0 24.5 20.1 21.6 17.1 -2.9 21.2 18.7 +2.6 BART _{small} 36.6 16.9 41.8 21.1 21.0 16.5 23.4 19.3 10.2 19.8 12.8 45.4 7.1 51.6 46.3 53.2 46.2 4.8 31.7 2.7 27.6 23.1 4.3 25.7 22.8 19.1 +2.0 PEGA _{base} 445 23.1 43.9 23.2 264 22.1 31.3 26.1 15.4 34.3 13.3 42.3 42.3 45.3 47.1 51.6 15.4 24.3 13.3 42.3 42.3 42.3 48.7 41.9 4.2.6 19.3 10.2 19.8 12.8 45.4 7.1 51.6 15.6 23.1 4.3 25.7 22.2 18.8 4.1 22.3 19.1 +2.0 PEGA _{base} 31.9 13.2 40.3 19.2 14.7 11.4 19.3 15.9 8.7 20.2 10.2 36.3 46.8 45.4 47.1 51.4 6.1 36.7 22.6 23.1 4.3 25.7 22.8 40.0 PEGA _{base} 31.9 13.2 40.3 19.2 14.7 11.4 19.3 15.9 8.7 20.2 10.2 36.3 46.8 45.4 47.1 34.9 41.9 4.2 41.9 4.1 20.1 16.7 0.6 16.4 13.9 13.9 45.7 40.3 40.3 45.1 45.1 45.1 45.1 45.1 45.1 45.1 45.1	3ART small 313 12.7 39.5 18.5 5.8 15.1 13.8 44.4 +8.9 52.1 46.7 56.1 49.6 +3.4 24.8 20.9 23.6 19.5 1.3 22.3 20.0 +1.7 56.1 BEGA base 45.2 23.6 43.3 22.8 23.7 12.5 33.2 13.3 37.8 +0.4 55.4 45.5 5.5 29.3 25.1 28.8 24.9 -0.4 25.4 55.4 Average 28.3 10.7 40.4 18.6 15.0 91.5 15.0 91.9 16.9 +0.9 16.1 13.8 +0.0 46.9 Stands 27.7 7.4 18.7 15.6 15.0 91.7 40.7 35.5 43.9 36.7 +2.2 7.5 61.1 13.8 40.9 25.4 51.3 32.7 14.6 90.9 16.9 40.9 23.7 12.5 32.7 45.3 7.6 11.1 16.9 16.1 15.4 15.4 15.4 15.6 16.9 16.9 16.9	10116-1	$BART_{base}$	31.3	13.5	40.6	20.2									51.9	46.8	55.1		+2.5										-2.1
$ \begin{array}{rrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrr$	EGA.base 452 23.6 43.3 22.8 23.7 12.5 33.2 13.3 37.8 +0.4 55.4 49.7 61.5 54.5 55.1 28.8 24.9 -0.4 25.4 21.8 23.7 12.5 33.2 13.3 37.8 +0.4 55.4 49.7 61.5 54.5 55.1 28.8 24.9 -0.4 25.4 21.8 21.4 50.1 46.9 10.9 10.1 13.8 +0.0 46.9 15.7 13.7 +0.1 15.9 10.1 13.8 +0.0 46.9 13.7 15.9 15.1 15.9 15.0 15.9 15.0 15.7		BART _{small}	31.3	12.7	39.5	18.5									52.2	46.7	56.1		+3.4									50.5	+3.8
Average 28.3 10.7 40.4 18.6 16.4 15.5 4.9 15.0 9.9 35.0 48.2 37.9 50.1 43.7 46.8 19.0 15.9 16.9 40.9 16.1 13.8 40.0 T5buse 27.9 7.4 42.4 18.7 2.4 0.0 9.2 15.0 16.4 4.9 22.7 $+7.5$ 40.7 35.5 43.9 36.7 $+2.2$ 7.5 6.1 13.3 12.1 $+5.9$ 8.2 5.2 $+5.1$ T5smult 14.8 1.2 31.8 11.8 4.0 2.4 10.8 25.3 26.1 $+11.8$ 25.3 20.8 32.4 28.5 $+7.4$ 11.6 9.6 13.7 12.4 $+2.5$ 3.7 $+0.9$ BART _{numl} 36.6 16.9 41.8 10.9 9.2 0.2 9.5 45.4 $+7.1$ 51.6 40.3 36.7 42.5 43.5 21.2 18.7	Average 28.3 10.7 40.4 18.6 16.4 15.5 4.9 15.0 9.9 35.0 48.7 50.1 43.7 46.8 19.0 15.9 16.9 40.9 16.1 13.8 40.0 46.9 T5buse 27.9 7.4 42.4 18.7 2.4 0.0 9.0 15.0 16.1 13.8 40.0 46.9 T5buse 27.9 7.4 42.4 18.7 2.4 0.0 9.5 45.1 57.1 32.7 T5buse 27.9 7.4 2.1 8.7 2.2 4.5 35.7 43.9 36.7 +2.2 7.5 6.1 13.3 12.1 45.9 82 5.2 45.1 32.7 Stand 14.8 1.2 31.8 11.8 4.0 2.4 40.7 35.5 43.1 2.0 23.7 40.9 23.7 49.0 46.9 23.7 47.6 49.0 23.7 49.0 46.9 23.7 49.0 46.9 23.7 49.1 23.7 49.0 46.9 23.7		PEGA.base	45.2	23.6	43.3	22.8									55.4	49.7	61.5		+5.5								55.4		+0.4
$ T5_{\text{base}} \left[\begin{array}{cccccccccccccccccccccccccccccccccccc$	T5base 279 7.4 42.4 187 2.4 0.8 8.1 6.1 3.0 16.4 4.9 22.7 +7.5 40.7 35.5 43.9 36.7 +2.2 7.5 6.1 13.3 12.1 +5.9 82 5.2 +5.1 32.7 T5small 14.8 1.2 31.8 11.8 4.0 2.4 10.9 9.2 0.2 3.5 26.1 +11.8 25.3 20.8 32.4 28.5 +7.4 11.6 9.6 13.7 12.4 +2.5 4.5 3.7 +0.9 23.7 ARTTass 35.8 17.5 41.4 21.2 19.5 15.3 22.9 18.9 14.8 30.4 16.5 45.1 51.6 43.1 2.0 23.7 12.0 12.4 +2.5 45.7 49.0 23.7 ARTTass 35.6 16.9 41.8 20.1 21.1 11.6 9.6 13.7 12.4 49.0 23.7 49.0 33.7 BGRAbase 44.5 23.1 43.3 32.3 46.2 <th></th> <th>Average</th> <th>28.3</th> <th>10.7</th> <th>40.4</th> <th>18.6</th> <th></th> <th></th> <th></th> <th></th> <th></th> <th></th> <th>35.0</th> <th></th> <th>42.4</th> <th>37.9</th> <th>50.1</th> <th></th> <th>+6.8</th> <th>19.0</th> <th></th> <th></th> <th></th> <th></th> <th></th> <th></th> <th></th> <th></th> <th>+4.1</th>		Average	28.3	10.7	40.4	18.6							35.0		42.4	37.9	50.1		+6.8	19.0									+4.1
$ T5_{\text{stuall}} \left[\begin{array}{cccccccccccccccccccccccccccccccccccc$	T5_small 14.8 1.2 31.8 11.8 4.0 2.4 10.9 9.2 0.2 3.5 26.1 +11.8 25.3 20.8 32.4 28.5 +7.4 11.6 9.6 13.7 12.4 +2.5 4.5 3.7 +0.9 23.7 ARTTasse 35.8 17.5 41.4 21.2 19.5 15.3 22.9 18.9 14.8 30.4 16.5 45.1 +5.4 52.6 45.8 53.1 -2.0 24.5 17.1 -2.9 21.2 18.7 +2.6 49.0 ARTTasse 35.6 16.9 41.8 21.1 21.0 16.5 23.4 19.3 10.2 19.8 12.8 45.4 +7.1 51.6 46.3 53.2 43.1 -2.0 24.5 13.1 27.7 27.6 29.1 14.2 23.1 43.9 25.7 28.7 43.6 49.0 58.7 45.5 45.5 45.6 45.6 45.7 25.6 45.7 25.6 45.7 45.6 49.0 58.7 43.5 45.7 45.6		$T5_{base}$	27.9	7.4	42.4	18.7							22.7		40.7	35.5	43.9		+2.2								32.7		-7.8
BART _{isse} 35.8 17.5 41.4 21.2 19.5 15.3 22.9 18.9 14.8 30.4 16.5 45.1 +5.4 52.6 46.8 52.2 43.1 -2.0 24.5 20.1 21.6 17.1 -2.9 21.2 18.7 +2.6 BART _{standl} 36.6 16.9 41.8 21.1 21.0 16.5 23.4 19.3 10.2 19.8 12.8 45.4 +7.1 51.6 46.3 53.2 46.2 +0.8 26.7 22.5 22.2 18.8 4.1 22.3 19.1 +2.0 PEGA _{base} 44.5 23.1 4.3 23.2 26.4 22.1 31.3 26.1 15.4 34.3 13.3 42.3 +2.4 57.4 52.0 61.9 55.1 +3.8 31.7 27.7 27.6 23.1 4.3 25.7 22.8 +0.0 Average 31.9 13.2 40.3 19.2 14.7 11.4 19.3 15.9 8.7 20.2 10.2 36.3 +6.8 45.5 40.3 48.7 41.9 +2.4 20.4 17.2 19.7 16.7 -0.6 16.4 13.9 +2.1	ART have 35.8 17.5 41.4 21.2 19.5 15.3 22.9 18.9 14.8 30.4 16.5 45.1 +5.4 52.6 46.8 52.2 43.1 -2.0 24.5 20.1 21.6 17.1 -2.9 21.2 18.7 +2.6 49.0 ART main 36.6 16.9 41.8 21.1 21.0 16.5 23.4 19.3 10.2 19.8 12.8 45.4 +7.1 51.6 46.3 53.2 46.2 +0.8 26.7 22.5 23.1 43.9 23.2 19.1 +2.0 47.8 41.9 47.8 47.8 47.1 27.6 29.7 12.8 47.1 20.4 47.8 47.8 47.8 47.8 47.8 47.8 47.8 47.8 47.8 47.8 47.8 47.8 47.8 47.8 47.8 47.8 47.8 47.1 47.8 47.8 47.1 27.7 27.6 23.1 43.3 47.8 47.8 47.8 47.8 47.8 47.8 47.8 47.8 47.8 47.8 47.8 <th></th> <td>$T5_{small}$</td> <td>14.8</td> <td>1.2</td> <td>31.8</td> <td>11.8</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td>26.1</td> <td>+11.8</td> <td></td> <td>20.8</td> <td>32.4</td> <td></td> <td>+7.4</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td>-1.9</td>		$T5_{small}$	14.8	1.2	31.8	11.8							26.1	+11.8		20.8	32.4		+7.4										-1.9
36.0 16.9 41.8 21.1 21.0 16.5 23.4 19.3 10.2 19.8 12.8 45.1 51.6 46.3 53.2 46.2 +0.8 26.7 22.5 22.3 19.1 22.3 19.1 +2.0 44.5 23.1 43.9 15.4 34.3 13.3 42.3 43.3 13.3 42.3 43.3 13.3 42.4 57.4 57.0 61.9 55.1 +3.8 31.7 27.7 27.6 23.1 4.3 25.7 22.8 +0.0 31.9 13.2 40.3 19.2 19.3 15.9 8.7 20.2 10.2 36.3 +6.8 45.5 40.3 48.7 41.9 +2.4 20.4 17.2 19.7 16.7 -0.6 16.4 13.9 +2.1 31.9 13.2 40.3 19.7 41.9 +2.4 20.4 19.7 16.7 -0.6 16.4 13.9 +2.1	ART small 36.6 16.9 41.8 21.1 21.0 16.5 23.4 19.3 10.2 19.8 12.8 45.4 +7.1 51.6 46.3 53.2 46.2 +0.8 26.7 22.5 22.3 19.1 +2.0 47.8 BEGAbase 44.5 23.1 43.9 23.2 26.4 23.1 34.3 31.3 26.1 15.4 34.3 13.3 42.3 42.4 57.4 52.0 61.9 55.1 +3.8 31.7 27.7 27.6 23.1 4.3 25.7 22.8 +0.0 58.7 Average 31.9 13.2 40.3 15.7 20.6 15.4 55.1 45.3 45.7 27.6 23.1 4.3 25.7 22.8 +0.0 58.7 Average 31.9 13.2 40.3 15.7 20.4 15.7 16.7 -0.6 16.4 13.9 +2.1 42.4 Average 31.9 13.2 40.3 15.7 20.4 15.7 16.7 -0.6 16.4 13.9 +2.1 42.	8-shot	$BART_{base}$	35.8	17.5	41.4	21.2							-	+5.4	52.6	46.8	52.2	43.1	-2.0										-4.3
44.5 23.1 43.9 23.2 26.4 22.1 31.3 26.1 15.4 34.3 42.0 61.9 55.1 +3.8 31.7 27.7 27.6 23.1 4.3 20.7 22.8 40.0 31.9 13.2 40.3 19.2 14.9 15.4 19.4 19.3 15.9 8.7 20.2 10.2 36.3 +6.8 45.5 40.3 48.7 41.9 +2.4 20.4 15.7 22.8 +0.0	FEGAbase 44.5 23.1 43.9 23.2 26.4 34.3 13.3 42.3 13.3 42.3 13.4 13.4 57.4 57.4 57.4 57.4 57.4 57.1 23.1 4.3 23.1 4.3 25.7 22.8 +0.0 58.7 Average 31.9 13.2 40.3 19.2 14.7 11.4 19.3 15.9 8.7 40.3 48.7 41.9 +2.4 20.4 17.2 19.7 16.7 0.6 16.4 13.9 +2.1 42.4		BART _{small}	36.6	16.9	41.8	21.1									51.6	46.3	53.2	·	+0.8										-4.6
31.9 13.2 40.3 19.2 14.7 11.4 19.3 15.9 8.7 20.2 10.2 36.3 +6.8 45.5 40.3 48.7 41.9 +2.4 20.4 17.2 19.7 16.7 -0.6 16.4 13.9 +2.1	Average 31.9 13.2 40.3 19.2 14.7 11.4 19.3 15.9 8.7 41.9 +2.4 20.4 17.2 19.7 16.7 -0.6 16.4 13.9 +2.1 42.4		PEGA.base	44.5	23.1	43.9	23.2									57.4	52.0	61.9		+3.8									51.9	+0.6
			Average	31.9	13.2	40.3	19.2									45.5	40.3	48.7		+2.4								42.4		-3.6

CLU3-2 allu 5 2 2 3 5 2 ż 5, ì 1 5, ן ά 2 . È ĵ 2 Table 9: Full experiment results. K-L: because they use shared few-shot tasks.