

Less for More: Enhanced Feedback-aligned Mixed LLMs for Molecule Caption Generation and Fine-Grained NLI Evaluation

Anonymous ACL submission

Abstract

Scientific language models drive research innovation but require extensive fine-tuning on large datasets. This work enhances such models by improving their inference and evaluation capabilities with minimal or no additional training. Focusing on molecule caption generation, we explore synergies between alignment fine-tuning and model merging in a cross-modal setup. We reveal intriguing insights into the behaviour and suitability of such methods while significantly surpassing state-of-the-art models. Moreover, we propose a novel atomic-level evaluation method leveraging off-the-shelf Natural Language Inference (NLI) models for use in the unseen chemical domain. Our experiments demonstrate that our evaluation operates at the right level of granularity, effectively handling multiple content units and subsentence reasoning, while widely adopted NLI methods consistently misalign with assessment criteria.

1 Introduction

AI in Chemistry is essential for developing scalable and cost-effective scientific solutions, such as pioneering drugs (Ferguson and Gray, 2018), advanced materials (Kippelen and Brédas, 2009), and improved chemical processes (Zhong et al., 2023). The vast search spaces in which these solutions reside make chemical language models crucial for accelerating scientific discovery (AI4Science and Quantum, 2023; Zhang et al., 2023). Recent trends have led to the use of multimodal models to learn molecular and linguistic representations, either in separate but coordinated spaces (Edwards et al., 2021, 2022; Liu et al.,

2023a), in a common space (Liu et al., 2023b), or through dual approaches (Luo et al., 2023; Christofidellis et al., 2023). These models often rely heavily on extensive supervised fine-tuning. However, merely increasing model size and data does not guarantee improvement (Tirumala et al., 2022; Xu et al., 2023). Thus we propose focusing on novel training methods.

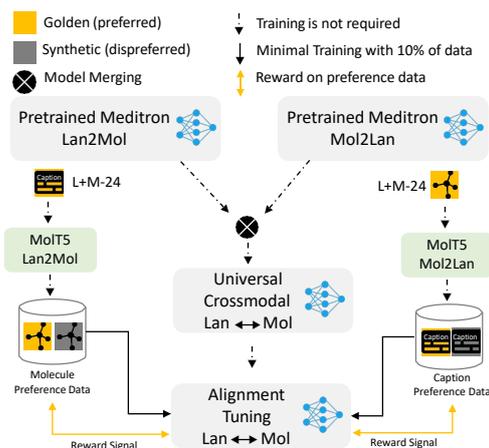


Figure 1: Overview of our proposed comprehensive solution to address key limitations in chemical LLMs, including extensive fine-tuning and out-of-distribution performance via model merging and alignment tuning with synthetic dispreferred data generated by MolT5.

Here we enhance molecule language models using minimal training by leveraging synergies between alignment fine-tuning (Ouyang et al., 2022) and model merging (Yang et al., 2024) in a cross-modal setup. Specifically, we focus on molecule-language translation, using as little as 10% of the training data (Edwards et al., 2024). Fig. 1 illustrates our comprehensive solution.

052	Model merging, a technique for fusing models	099
053	fine-tuned on different tasks, builds a versatile	100
054	model without needing the original training data	101
055	or expensive computation. This method has been	102
056	quickly adopted in foundation models and Large	103
057	Language Models (LLMs) (Yang et al., 2024).	104
058	We extend this concept to a crossmodal setting by	105
059	merging per-task pretrained molecule language	106
060	models (see Fig. 1), deploying both weight- and	107
061	subspace-based techniques to obtain universal	108
062	models (§ 3.2.1).	109
063	For fine-tuning alignment, we focus on Re-	110
064	inforcement Learning from Human Feedback	111
065	(RLHF)(Stiennon et al., 2020) to align the uni-	112
066	versal models. Although alignment has typically	113
067	been used to calibrate LLM behaviour (Askell	114
068	et al., 2021), we hypothesise that it can also accel-	115
069	erate learning in crossmodal spaces by rewarding	116
070	preferred over dispreferred outputs, thus improv-	117
071	ing inference with minimal training data. We fo-	118
072	cus on optimisation algorithms using closed-form	119
073	losses on offline preferences, such as Direct Pref-	120
074	erence Optimisation (DPO) (Rafailov et al., 2024),	121
075	Contrastive Preference Optimisation (CPO) (Xu	122
076	et al., 2024), and Kahneman-Tversky Optimisa-	123
077	tion (KTO) (Ethayarajh et al., 2024). We incorpo-	124
078	rate golden data as human preferences and dispre-	125
079	ferred synthetic outputs generated by proprietary	126
080	models into the reward signal (see Fig. 1).	127
081	We evaluate our models on out-of-distribution	128
082	data using established statistical-based met-	129
083	rics (Sets, 2022; Edwards et al., 2022). Addition-	130
084	ally, we use Natural Language Inference (NLI)	131
085	models to assess generated text within the chemi-	
086	cal domain. However, we argue that off-the-shelf	
087	NLI models are suboptimal for several reasons: a)	
088	they are trained on relatively short texts (Williams	
089	et al., 2018), while generated text may aggregate	
090	multiple content units that partially overlap with	
091	different sentences in the reference text (Nenkova	
092	et al., 2007); b) they are limited by the data they	
093	were trained on, making them unreliable for un-	
094	seen domains (McIntosh et al., 2024); and c) they	
095	lack subsentence inference, hindering their ability	
096	to handle reordered content in generated text (see	
097	Fig. 3). Thus we propose a novel atomic-level	
098	cross-NLI approach that addresses these issues.	
	By decomposing reference and generated texts	
	into atomic premises and hypotheses using an	
	LLM, we calculate probability distributions of	
	contradiction and entailment via an NLI model	
	and finally apply row-wise operations to obtain	
	novel hallucination and coverage metrics (§3.3).	
	Our findings and contributions are as follows:	
	• Extensive training doesn’t guarantee better	
	models. Models trained on large benchmark	
	datasets exhibit memorisation effects, with per-	
	formance dropping by 50% to 100% on out-of-	
	distribution data (§ 4.2.1).	
	• Alignment fine-tuning is not a panacea. Our	
	experiments reveal that not all fine-tuning ap-	
	proaches applicable to heavily trained models	
	are effective with minimal training (§ 4.2.1).	
	• Effective alignment methods balance struc-	
	tured learning and generalisation. Of the	
	alignment fine-tuning methods, only CPO man-	
	aged both crossmodal agnostic and minimal	
	training effectively (§ 4.2.1).	
	• Model merging addresses inherent limita-	
	tions in alignment fine-tuning. It improves	
	performance with minimal training, reduces de-	
	pendence on human-labeled data, and provides	
	a scalable, cost-effective alignment method for	
	LLMs. (§ 4.2.2).	
	• Our novel atomic-level cross-NLI evalua-	
	tion reveals intriguing insights about perfor-	
	mance interpretability and effectively han-	
	dles multiple content units in text. By con-	
	trast, widely adopted NLI methods consistently	
	misalign with assessment criteria (§ 4.2.3).	
	2 Related Work	132
	2.1 LLMs for Chemistry	133
	Existing approaches for LLMs in the chemi-	134
	cal domain typically rely on costly pretraining	135
	with large unimodal datasets for reaction predic-	136
	tion and retrosynthesis (Schwaller et al., 2019;	137
	Vaucher et al., 2020), or task-specific fine-tuning	138
	for language-molecule learning (Edwards et al.,	139
	2021, 2022, 2024) and molecule editing (Liu	140
	et al., 2023a; Fang et al., 2023). Other meth-	141
	ods focus on multitask learning, which requires	142
	resource-intensive pretraining and large multitask	143

144	datasets (Lu and Zhang, 2022; Ross et al., 2022;	as the premise (sentence-level NLI)(Nie et al.,	188
145	Christofidellis et al., 2023; Zhang et al., 2024). In	2019b; Laban et al., 2022), or use the entire refer-	189
146	contrast, we investigate synergies between fine-	ence as the premise(Dziri et al., 2022; Honovich	190
147	tuning alignment and model merging to enhance	et al., 2022), which can be inefficient for long	191
148	molecule language models with minimal training.	texts (Schuster et al., 2022). Context-level NLI	192
149	2.2 Model Merging	addresses this by retrieving relevant sentences to	193
150	Existing model merging techniques can be	create a short context (Nie et al., 2019a; Schus-	194
151	broadly categorised into weight-based, subspace-	ter et al., 2022; Kamoi et al., 2023), but lacks	195
152	based, and routing-based approaches. Weight-	sufficient granularity (Nenkova et al., 2007). We	196
153	based methods often use optimisation algo-	propose a novel atomic-level NLI evaluation for	197
154	rithms (Yang et al., 2023; Akiba et al., 2024)	the chemical domain to address these limitations.	198
155	or geometric interpolations (Zhou et al., 2024;		
156	Goddard et al., 2024) to determine optimal task	3 Methodology	199
157	vector coefficients. Subspace-based methods in-	3.1 Task Definition	200
158	volve pruning (Yadav et al., 2023; Yu et al.,	Let (x, y) represent a pair of source and target	201
159	2024) or masking (Wang et al., 2024) to remove	sequences mapped to the X and Y spaces, re-	202
160	insignificant parameters, reducing task interfer-	spectively. We cast molecule caption genera-	203
161	ence. Routing-based methods combine models	tion (MoCG) as a crossmodal alignment task	204
162	adaptively during inference based on specific in-	that operates on offline preference data $\mathcal{D} =$	205
163	put (Muqeeth et al., 2023; Tang et al., 2024).	$\{x^{(i)}, y_w^{(i)}, y_l^{(i)}\}_{i=1}^N$, where x is the input, and y_w	206
164	We experiment with weight- and subspace-based	and y_l are the preferred and dispreferred outputs,	207
165	merging in a crossmodal context.	respectively, with N being the total number of	208
166	2.3 Aligning LLMs	pairs in \mathcal{D} . The goal is to learn an optimal func-	209
167	LLM alignment methods can be divided into test-	tion $f : X \leftrightarrow Y$ via a model π_θ parameterised	210
168	time and fine-tuning approaches. Test-time align-	by θ . We coordinate the molecule and caption	211
169	ment techniques, such as prompt engineering and	generation tasks via instruction modelling ¹ .	212
170	guided decoding (Khanov et al., 2024; Huang	3.2 Aligned Mixed Molecule Language	213
171	et al., 2024), adjust LLMs without changing their	Models	214
172	weights, but depend on the original model’s per-	This section elaborates on how we obtain aligned	215
173	formance. Fine-tuning methods, like RLHF (Sti-	universal molecule language models.	216
174	ennon et al., 2020; Ouyang et al., 2022), are effec-		
175	tive but complex, requiring model retraining and	3.2.1 Universal Models via Model Merging	217
176	continuous sampling. DPO (Rafailov et al., 2024)	Let τ_1 and τ_2 represent task vectors ² from pre-	218
177	simplifies RLHF by directly optimizing PPO’s ob-	trained molecule and caption generation models.	219
178	jective, while CPO (Xu et al., 2024) improves effi-	Our goal is to obtain a multitasking cross-modal	220
179	ciency by using a uniform reference model. Other	model $\Theta^{(merge)}$ without accessing training data	221
180	methods leverage SFT for optimizing RLHF man-	by exploring weight-based and subspace-based	222
181	agement and parameter tuning (Ethayarajh et al.,	merging techniques. Fig. 2 illustrates the process.	223
182	2024; Meng et al., 2024). Here, we explore align-	Specifically, we experiment with model merging	224
183	ment fine-tuning in a crossmodal setup.	approaches that inherently manage conflicts and	225
184	2.4 NLI-based Evaluation	mitigate modality dominance or instability when	226
185	NLI models determine the relationship between	integrating modality-specific information using	227
186	a <i>premise</i> and a <i>hypothesis</i> . Existing approaches		
187	either identify a sentence in the reference text		

¹Instructions can be found in Appx. F.

²A task vector τ represents the model’s parameters $\Theta^{(t)}$ fine-tuned for task t (Ilharco et al., 2022).

228
229

off-the-shelf LLMs, ensuring that neither modality overshadows the other.

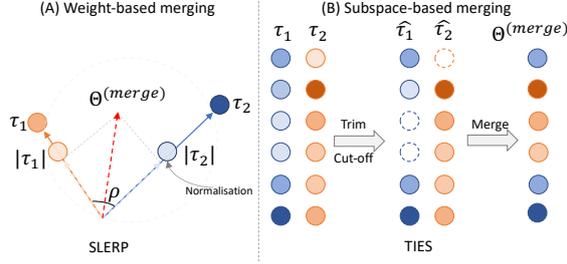


Figure 2: Model merging techniques for obtaining universal models. (A) Weight-based merging via spherical interpolation. (B) Subspace-based merging by pruning and merging parameter magnitudes. τ_1 and τ_2 are task vectors obtained from pretrained molecule and caption generation models, respectively.

230
231
232
233
234
235
236
237
238
239

Weight-based model merging: We experiment with SLERP (Goddard et al., 2024), which applies spherical interpolation to fuse model parameters. The goal is to find optimal coefficients λ_1 and λ_2 so that the merged model $\Theta^{(merge)} = \lambda_1 \tau_1 + \lambda_2 \tau_2$ retains the capabilities of the independent models. The coefficients are given by $\frac{\sin((1-\lambda)\cdot\rho)}{\sin(\rho)}$ and $\frac{\sin(\lambda\cdot\rho)}{\sin(\rho)}$, respectively, where $\rho = \arccos\left(\frac{\tau_1 \cdot \tau_2}{|\tau_1| |\tau_2|}\right)$ is the angle between the task vectors, and λ is the merging coefficient.

240
241
242
243
244
245
246
247
248
249

Subspace-based model merging: We utilise TIES (Yadav et al., 2023) to prune the task vectors τ_1 and τ_2 , retaining the top 20% parameters, resulting in refined vectors $\hat{\tau}_1$ and $\hat{\tau}_2$ (see Fig. 2 (B)). We then fuse the vectors via Task Arithmetic (Ilharco et al., 2022) to obtain the merged model as $\Theta^{(merge)} = \frac{1}{2} \sum_{i=1}^2 \hat{\tau}_i$. During the merging process, conflicts arising from differing signs in the parameters p are resolved by aligning the pruned vectors as follows:

250

$$\text{Align}(\hat{\tau}_1^p, \hat{\tau}_2^p) = \begin{cases} \hat{\tau}_1^p & \text{if } |\hat{\tau}_1^p| > |\hat{\tau}_2^p| \\ \hat{\tau}_2^p & \text{if } |\hat{\tau}_2^p| \geq |\hat{\tau}_1^p| \end{cases} \quad (1)$$

251
252
253
254

3.2.2 Crossmodal Alignment Fine-tuning

Let π_{ref} be the reference policy (i.e., the universal model from model merging), π_θ the policy model being trained, parameterised by θ , and

$\mathcal{D} = \{x^{(i)}, y_w^{(i)}, y_l^{(i)}\}$ the offline preference data. Our goal is to learn effective crossmodals for the MoCG task with minimal training via alignment fine-tuning. We experiment with different optimizations that differ substantially in how they learn a reward signal, as overviewed in Table 1.

255
256
257
258
259
260

- **SFT** minimises the difference between generated output z and target y_w by optimising model π_θ through negative log-likelihood (Eq. 2).
- **DPO** (Rafailov et al., 2024) enhances crossmodal translations using an offline preference dataset \mathcal{D} . It aligns model π_θ by maximising the likelihood of preference data, with reference model π_{ref} , Sigmoid function σ , and hyperparameter β (Eq. 3).
- **CPO** (Xu et al., 2024) reduces reliance on high-quality data by avoiding suboptimal translations. It modifies Eq. 3 using a uniform reference model, ensuring equal likelihood for all outputs. A behaviour cloning (BC) regulariser is injected to reflect uniform output matching, with an additional SFT term in the final loss (Eq. 4).
- **KTO** (Ethayarajh et al., 2024) utilises non-paired preference data $\mathcal{D} = \{x^{(i)}, y^{(i)}, \lambda^{(i)}\}$ where λ denotes the desirability of y . The loss is computed from the generated output z in relation to a reference z_{ref} and λ (Eq. 5).

261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276

Method	Optimisation Objective
SFT	$\min_{\theta} -\log \pi_{\theta}(y_w x) \quad (2)$
DPO	$\log \sigma \left(\beta \log \frac{\pi_{\theta}(y_w x)}{\pi_{ref}(y_w x)} - \beta \log \frac{\pi_{\theta}(y_l x)}{\pi_{ref}(y_l x)} \right) \quad (3)$
CPO	$\min_{\theta} \log \sigma \left(\beta \log \pi_{\theta}(y_w x) - \beta \log \pi_{\theta}(y_l x) \right) - \log \pi_{\theta}(y_w x)$ s.t. $\mathbb{E}_{(x,y_w) \sim \mathcal{D}} \left[\text{KL}(\pi_w(y_w x) \pi_{\theta}(y_w x)) \right] < \epsilon \quad (4)$
KTO	$-\lambda_w \sigma \left(\beta \log \frac{\pi_{\theta}(y_w x)}{\pi_{ref}(y_w x)} - z_{ref} \right) + \lambda_l \sigma \left(z_{ref} - \beta \log \frac{\pi_{\theta}(y_l x)}{\pi_{ref}(y_l x)} \right)$ where $z_{ref} = \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\beta \text{KL}(\pi_{\theta}(y x) \pi_{ref}(y x)) \right] \quad (5)$

Table 1: Alignment fine-tuning algorithms for the MoCG task given preference data $\mathcal{D} = \{x, y_w, y_l\}$.

- **KTO** (Ethayarajh et al., 2024) utilises non-paired preference data $\mathcal{D} = \{x^{(i)}, y^{(i)}, \lambda^{(i)}\}$ where λ denotes the desirability of y . The loss is computed from the generated output z in relation to a reference z_{ref} and λ (Eq. 5).

277
278
279
280
281

3.3 Atomic-level Cross-NLI Evaluation

Our aim is to develop a method that operates at the right level of granularity, precisely capturing small distinctions and subtle nuances in captions, ensuring reliable evaluation. Atomic-level cross-NLI evaluation uses a LLM and an NLI model to assess relationships between generated and reference captions. The process begins with an LLM (Touvron et al., 2023) decomposing a (reference, generated) pair into atomic premises $\{P_i\}_{i=1}^N$ and hypotheses $\{H_j\}_{j=1}^L$, where each atomic unit conveys a single piece of information (see Appx. E). An NLI model (He et al., 2020) then constructs probabilistic distributions of entailment and contradiction by considering all possible combinations of premises and hypotheses. Finally, pooling operators match atomic hypotheses and premises in terms of both factual correctness, i.e., *hallucination*, and completeness, i.e., *coverage*. Fig. 3 illustrates this process.

Hallucination we define here as the introduction of information not present in the reference text. Given $\{(P_i, H_j)\}$, the NLI model constructs a contradiction probability distribution for each atomic hypothesis against all premises, such as $p_{j,i} = (C_{j,i}|P_i, H_j)$. This results in an $M_{L \times N}$ matrix of contradiction probabilities $C_{j,i}$ (see Fig. 3). To measure hallucination, we apply min row-wise pooling and average the matching probabilities to compute the score by the formula:

$$Hallucination = \frac{1}{L} \sum_{j=1}^L \min_i C_{j,i} \quad (6)$$

Coverage we define as atomic unit recall, representing how much reference information is present in the generated text. Unlike hallucination, here generated text forms the atomic premises (P_j) and the reference text the hypotheses (H_i). The NLI model constructs an entailment probability distribution for each H_i against all P_j , such that $p_{i,j} = (E_{i,j}|P_j, H_i)$, resulting in an $M_{N \times L}$ matrix of entailment probabilities $E_{i,j}$. To measure coverage, we apply max row-wise pooling and average the matching probabilities to compute the score given by the formula:

$$Coverage = \frac{1}{N} \sum_{i=1}^N \max_j E_{i,j} \quad (7)$$

4 Experiments

4.1 Experimental Setup

Data: We conduct experiments training Meditron (Chen et al., 2023) on the benchmark L+M-24 (Edwards et al., 2024) dataset, using only 10% of the data for training, and evaluate on out-of-distribution data (see Appx. D for details). For alignment fine-tuning, we create synthetic dispreferred outputs generated by MolT5 (Edwards et al., 2022). In practice, this involves feeding MolT5 with inputs from the 10% subset of L+M-24 used in our experiments, generating outputs, and then using these outputs as dispreferred samples (see Fig. 1). Our training, validation, and test sets contain approximately 12.7k, 3.4k, and 3k samples.

Baselines: We selected established baselines based on their relevance to our hypotheses, enabling comparison with models trained on fully (i.e., Chem-LLM (Zhang et al., 2024)) and partially (i.e., TxtChem-T5 (Christofidellis et al., 2023)) out-of-distribution data, as well as in-distribution data (Meditron (Chen et al., 2023)). In this context, TxtChem-T5 and Chem-LLM are evaluated in a zero-shot setting. For more details about the baselines, please refer to Appx. G. Lastly, we fine-tune Meditron with *SFT* using only 10% of the training data. We leave all the implementation details in Appx. J.

Evaluation: When evaluating the performance of both baselines and our models, we employ established statistical metrics (see Appendix H), in addition to our atomic-level cross-NLI evaluation method (§ 3.3). For our proposed evaluation, we assess the robustness of different NLI methods by measuring the entropy of textual entailment between generated outputs from high and low performance models in association with linguistic ones derived by bioinformatic databases curated by humans. Specifically, we compare our atomic-level NLI approach with leading ones, including

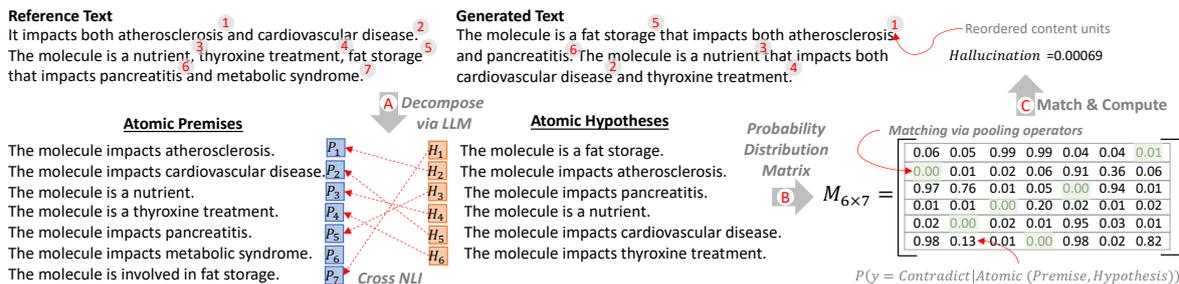


Figure 3: The process of atomic-level cross-NLI evaluation when measuring the level of hallucination.

367 *full NLI*, which treats entire premises and hypothe- 402
368 ses as single units, and *sentence-level NLI* (Laban 403
369 et al., 2022), which evaluates chunks in text. 404

370 4.2 Experimental Results 405

371 4.2.1 Aligning Molecule-Language Modals 407 372 with Minimal Training

373 We first present results for molecule language 408
374 models with minimal alignment fine-tuning, ini- 409
375 tialising pretrained weights from molecule gener- 410
376 ation rather than deploying model merging (see 411
377 Appx. J for details). Tables 2 and 3 summarise 412
378 experimental results. Generally, benchmarking 413
379 models trained on extensive data with SFT exhibit 414
380 memorisation effects, with performance dropping 415
381 by 50% to 100% compared to reported results, 416
382 when evaluated on out-of-distribution data. 417

383 Our experiments show that not all alignment 418
384 optimisations are effective in the minimal training 419
385 setting. Both DPO and KTO show zero perfor- 420
386 mance in caption generation when models are 421
387 initialised with crossmodal weights unrelated to 422
388 the task (see Table 2). However, performance 423
389 improves significantly when the crossmodals are 424
390 known (see Table 3). In molecule generation, 425
391 DPO achieves up to 42% better performance than 426
392 Meditron, trained on the full dataset, while KTO 427
393 still performs poorly, likely due to overfitting (see 428
394 Appx. I). 429

395 By contrast, CPO effectively handles both the 430
396 crossmodal agnostic and minimal training set- 431
397 tings, outperforming Meditron by up to 20% in 432
398 caption generation and 42% in molecule gener- 433
399 ation. This is likely due to its inherent ability 434
400 to balance structured learning and generalisation. 435
401 It aligns with preferred data through behaviour

402 cloning and SFT, which encourage the model 403
404 to mimic expert behaviour while reducing bias 405
406 and suboptimal outcomes via a uniform reference 407
408 model that assigns equal likelihood to all possible 409
410 outputs. 411

412 4.2.2 Alignment with Model Merging 417

418 Tables 4 and 5 summarise the experimental results 419
420 when we incorporate model merging in alignment 421
422 fine-tuning while keeping the training data the 423
424 same. Combining DPO with molecule and cap- 425
426 tion crossmodals via TIES improves caption gen- 426
427 eration (see $\Delta_{DPOvsTIES+DPO}$ in Table 4) but 427
428 leads to significant performance loss in molecule 428
429 generation (see $\Delta_{DPOvsTIES+DPO}$ in Table 5). 429
430 Conversely, fusing CPO with crossmodals via 430
431 SLERP significantly boosts performance in cap- 431
432 tion generation (see $\Delta_{CPOvsSLERP+CPO}$ in Ta- 432
433 ble 4) while having minimal impact on molecule 433
434 generation (see $\Delta_{CPOvsSLERP+CPO}$ in Table 5), 434
435 demonstrating overall gains compared to Med- 435
436 itron trained on the full dataset. 436

437 Overall, our experiments show that model 438
439 merging can effectively address key limitations 439
440 in alignment fine-tuning. By fusing pretrained 440
441 models, one can enhance performance with mini- 441
442 mal training, reducing reliance on human-labelled 442
443 data, lowering training costs, minimising human 443
444 bias, and improving generalisation. Examples of 444
445 caption and molecule generation are provided in 445
446 Appx. K. We leave further ablation experimental 446
447 studies in Appx. A. 447

448 4.2.3 Atomic-level Cross-NLI Evaluation 448

449 Atomic-level NLI revealed intriguing insights re- 450
451 garding performance interpretation. Fig. 4 shows 451
452 453

Method	Blue-2 \uparrow	Blue-4 \uparrow	Rouge-1 \uparrow	Rouge-2 \uparrow	Rouge-L \uparrow	METEOR \uparrow
TxtChem-T5 (Christofidellis et al., 2023)	0.08	0.09	0.19	0.06	0.17	0.16
Chem-LLM (Zhang et al., 2024)	0.03	0.00	0.11	0.02	0.09	0.14
Meditron (Chen et al., 2023)	0.42	0.30	0.63	0.47	0.49	0.54
SFT §4.1	0.37	0.26	0.55	0.40	0.39	0.61
DPO (Rafailov et al., 2024)	0.00	0.00	0.00	0.00	0.00	0.00
CPO (Xu et al., 2024)	0.62	0.45	0.68	0.50	0.48	0.62
KTO (Ethayarajh et al., 2024)	0.00	0.00	0.00	0.00	0.00	0.00
$\Delta_{CPOvsMED}$	+20%	+19%	+5%	+3%	-1%	+8%

Table 2: Alignment fine-tuning results for caption generation on 3k unseen pairs. Arrows next to metrics denote value increase with performance gains. Best results are in bold. $\Delta_{CPOvsMED}$ is the performance gain of our best model, trained on 10% of the data, compared to Meditron trained on the entire dataset.

Method	BLEU \uparrow	Levenshtein \downarrow	MACCS FTS \uparrow	RDKit FTS \uparrow	Morgan FTS \uparrow	FCD \downarrow	Validity \uparrow
TxtChem-T5	0.18	133.29	0.21	0.10	0.03	37.67	0.58
Chem-LLM	0.04	732.74	0.00	0.00	0.00	59.44	0.19
Meditron	0.43	66.16	0.35	0.29	0.19	13.64	0.57
SFT	0.30	186.99	0.70	0.62	0.41	11.14	0.98
DPO	0.72	42.40	0.77	0.69	0.49	10.47	0.99
CPO	0.71	42.65	0.77	0.70	0.48	4.19	1.00
KTO	0.23	294.63	0.03	0.03	0.02	32.64	0.06
$\Delta_{CPOvsMED}$	+29%	-23.76%	+42%	+41%	+30%	-9.45%	+41%

Table 3: Alignment fine-tuning results for molecule generation on 3k unseen pairs. Arrows next to metrics indicate whether higher or lower values denote better performance. Best results are highlighted in bold. $\Delta_{CPOvsMED}$ represents the performance gain of our best model compared to Meditron trained on the entire dataset.

436 assessment score distributions from our proposed
437 evaluation method, comparing our top models
438 against Meditron trained on the entire dataset. All
439 models exhibit low hallucination, likely due to the
440 narrow, well-defined topics that enable factually
441 correct captions without unrelated information.
442 However, our models excel in coverage, gener-
443 ating more comprehensive captions, with perfor-
444 mance increasing to 69% compared to Meditron’s
445 51% (Fig. 4 (B)). Examples of insights captured
446 by our proposed evaluation are in Appx. L.

447 We also evaluated the robustness of our pro-
448 posed NLI evaluation method against leading ap-
449 proaches by measuring the entropy of textual
450 entailment between human-curated texts (i.e.,
451 gold labels) and outputs generated by our top-
452 performing model, CPO+SLERP (preferred), ver-
453 sus those from a low-performing model, Med-
454 itron (dispreferred). Ideally, all NLI methods
455 should favour preferred outputs over dispreferred
456 ones. However, we observed that both the full and

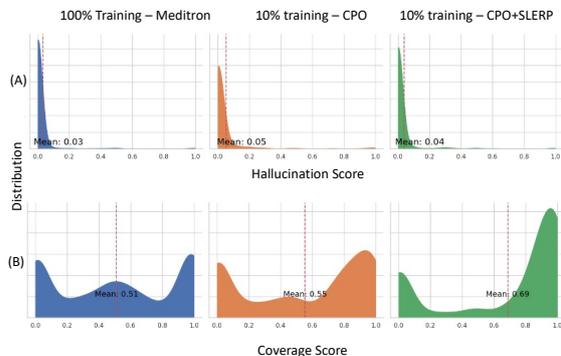


Figure 4: Score distributions from our atomic-level cross-NLI evaluation comparing (A) hallucination and (B) coverage between our top models and Meditron.

457 sentence-level NLI methods misclassify preferred
458 captions as non-entailment and dispreferred
459 captions as entailment (see Fig. 5 (B)-(D)). By con-
460 trast, atomic-level cross-NLI accurately favours
461 preferred captions, assigning higher scores to
462 certain cases (Fig. 5 (A)). Additionally, Kull-

Fusion	Method	Blue-2 \uparrow	Blue-4 \uparrow	Rouge-1 \uparrow	Rouge-2 \uparrow	Rouge-L \uparrow	METEOR \uparrow
TIES (Yadav et al., 2023)	DPO	0.74	0.53	0.74	0.54	0.51	0.70
	CPO	0.74	0.54	0.76	0.57	0.53	0.72
SLERP (Goddard et al., 2024)	DPO	0.00	0.00	0.02	0.01	0.00	0.00
	CPO	0.73	0.53	0.76	0.56	0.53	0.71
$\Delta_{DPOvsTIES+DPO}$		+74%	+53%	+74%	+54%	+51%	+70%
$\Delta_{CPOvsSLERP+CPO}$		+11%	+8%	+8%	+6%	+5%	+9%
$\Delta_{MEDvsSLERP+CPO}$		+31%	+28%	+13%	+9%	+4%	+17%

Table 4: Model merging and alignment fine-tuning results for caption generation. $\Delta_{DPOvsTIES+DPO}$, $\Delta_{CPOvsSLERP+CPO}$, and $\Delta_{MEDvsSLERP+CPO}$ measure performance gains of the best-combined approaches compared to the vanilla crossmodal setting of *DPO*, *CPO*, and the benchmark *Meditron*, as reported in Table 2.

Fusion	Method	BLEU \uparrow	Levenshtein \downarrow	MACCS FTS \uparrow	RDK FTS \uparrow	Morgan FTS \uparrow	FCD \downarrow	Validity \uparrow
TIES	DPO	0.32	93.18	0.31	0.22	0.19	19.80	0.42
	CPO	0.68	46.91	0.72	0.65	0.45	24.50	0.94
SLERP	DPO	0.72	43.85	0.77	0.70	0.51	10.35	0.98
	CPO	0.71	44.01	0.73	0.66	0.45	11.22	0.95
$\Delta_{DPOvsTIES+DPO}$		-40%	+51%	-46%	-47%	-30%	+7.33%	+58%
$\Delta_{CPOvsSLERP+CPO}$		0%	+1.36%	-4%	-4%	-3%	+5%	-4%
$\Delta_{MEDvsSLERP+CPO}$		+29%	-22.40%	+38%	+37%	+27%	-4.45%	+37%

Table 5: Model merging and alignment fine-tuning results for molecule generation. $\Delta_{DPOvsTIES+DPO}$, $\Delta_{CPOvsSLERP+CPO}$, and $\Delta_{MEDvsSLERP+CPO}$ measure performance gains of the best-combined approaches from the vanilla crossmodal setting of *DPO*, *CPO*, and the benchmark *Meditron*, as reported in Table 2.

463
464
465
466
467
468

back–Leibler divergence shows that atomic-level NLI offers better discrimination, achieving a divergence score of 0.54 compared to 0.12–0.17 for other methods, demonstrating its effectiveness in distinguishing the quality of generated captions. We leave further ablation analysis in Appx. B.

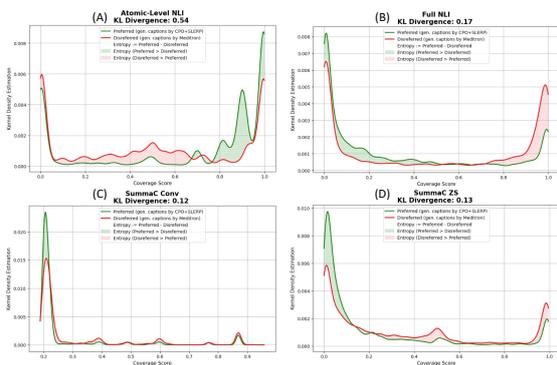


Figure 5: Relative entropy in coverage scores for preferred vs. dispreferred generated captions across atomic-level (A), full (B), and sentence-level (C & D) NLI approaches.

5 Conclusion

469

In this work, we address limitations of scientific language models that rely on extensive training. Focusing on molecule caption generation, we propose synergies between model merging and alignment fine-tuning with minimal training to enhance chemical language models. Our experiments show that while alignment fine-tuning performs poorly, incorporating model merging significantly outperforms extensively trained models on out-of-distribution data, offering a cost-effective approach that relies less on human-labelled data. Furthermore, we propose an atomic-level cross-NLI evaluation to overcome limitations of widely used NLI evaluation methods, which lack appropriate granularity. Our method provides valuable insight into performance interpretability and effectively handles multiple content units, where existing NLI methods consistently misalign with assessment criteria.

470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488

489	Limitations	
490	In this work, we employ weight-based and	535
491	subspace-based merging methods to create uni-	536
492	versal models for the MoCG task, facilitating	537
493	alignment fine-tuning in a training setting with	538
494	minimal data. However, both are static merging	539
495	methods. This means that the merged model re-	540
496	main the same for all samples or tasks. Given that	541
497	there are differences between input samples/tasks,	542
498	the models' ability may vary when processing	543
499	different samples/tasks. In the future, we aim to	544
500	investigate dynamically merging models (or sub-	545
501	sets of layers) based on the samples/tasks during	546
502	the inference phase (Kang et al., 2024).	
503	We also propose an atomic-level NLI evalua-	547
504	tion method that successfully handles multiple	548
505	content units, offering valuable insights into per-	549
506	formance interpretability for caption generation,	550
507	where widely adopted NLI methods consistently	551
508	misalign with assessment criteria. However, de-	552
509	composing text into atomic units can be challeng-	553
510	ing for other tasks involving complex or lengthy	554
511	text. While this method captures nuanced con-	555
512	tent, there is a risk of over-fragmentation, which	556
513	may lead to a loss of context or coherence in	557
514	evaluation. Additionally, the effectiveness of this	558
515	approach relies heavily on the LLM for decompo-	559
516	sition and the NLI model for entailment and con-	560
517	tradiction assessment. If either model struggles	561
518	with domain-specific content (e.g., highly techni-	562
519	cal language), the evaluation could yield inaccur-	563
520	ate or biased results. Furthermore, if generated	564
521	texts introduce valid but creative or non-standard	565
522	content, this approach may penalise them by clas-	566
523	sifying such deviations as contradictions or hallu-	567
524	cinations, even when they provide accurate infor-	568
525	mation. Future work will need to address these	569
526	limitations across various domains.	570
527	Finally, the proposed methods in this work are	571
528	tailored specifically for the chemical domain, fo-	
529	cusing on tasks like molecule caption generation.	
530	While these techniques—such as model merging	
531	and alignment fine-tuning—show promising re-	
532	sults within this context, their ability to generalise	
533	to other domains or scientific fields is uncertain.	
534	Different domains may have distinct data struc-	
	tures, tasks, and requirements, which might not	535
	align well with the crossmodal setup used here.	536
	For instance, a method optimised for chemical	537
	language and molecular structures may not work	538
	as effectively in domains like physics or biology,	539
	where the types of entities and relationships differ	540
	significantly. This potential lack of generalisation	541
	highlights the need for future research to explore	542
	the applicability of the proposed approaches in	543
	diverse scientific domains beyond chemistry, aim-	544
	ing to adapt and validate the methods for varying	545
	data structures and task requirements.	546
	Ethical Considerations	547
	The potential for generating misleading or incor-	548
	rect information poses significant ethical consid-	549
	erations in this work, particularly given the sci-	550
	entific context in which the language models are	551
	applied. If the models produce inaccurate cap-	552
	tions or misrepresent molecular characteristics, it	553
	could lead to erroneous conclusions in research	554
	and applications that rely on these outputs. This	555
	risk is particularly critical in fields like chemistry,	556
	where precise data interpretation is vital for safety,	557
	compliance, and advancing scientific knowledge.	558
	Furthermore, the reliance on automated evalua-	559
	tions may not adequately catch nuanced errors	560
	that human experts would recognise, potentially	561
	allowing flawed outputs to go unchecked. There-	562
	fore, ensuring that the models maintain a high	563
	standard of accuracy and reliability is essential	564
	to prevent the dissemination of misinformation,	565
	which could undermine trust in automated sys-	566
	tems and hinder scientific progress. Addressing	567
	these ethical concerns requires implementing ro-	568
	burst validation mechanisms and continuously in-	569
	volving domain experts in the evaluation process	570
	to ensure the integrity of the generated content.	571
	References	572
	Microsoft Research AI4Science and Microsoft Azure	573
	Quantum. 2023. The impact of large language mod-	574
	els on scientific discovery: a preliminary study us-	575
	ing gpt-4. <i>arXiv preprint arXiv:2311.07361</i> .	576
	Takuya Akiba, Makoto Shing, Yujin Tang, Qi Sun,	577
	and David Ha. 2024. Evolutionary optimiza-	578

579	tion of model merging recipes. <i>arXiv preprint arXiv:2403.13187</i> .	629
580		630
581	Amanda Askeff, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. <i>arXiv preprint arXiv:2112.00861</i> .	631
582		632
583		633
584		634
585		635
586	Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. <i>Biometrika</i> , 39(3/4):324–345.	636
587		637
588		638
589		639
590	Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. <i>arXiv preprint arXiv:2311.16079</i> .	640
591		641
592		642
593		643
594		644
595		645
596		646
597	Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, and Matteo Manica. 2023. Unifying molecular and textual representations via multi-task language modelling. In <i>International Conference on Machine Learning</i> , pages 6140–6157. PMLR.	647
598		648
599		649
600		650
601		651
602		652
603	Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. <i>Advances in Neural Information Processing Systems</i> , 36.	653
604		654
605		655
606		656
607	Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2022. Evaluating attribution in dialogue systems: The begin benchmark. <i>Transactions of the Association for Computational Linguistics</i> , 10:1066–1083.	657
608		658
609		659
610		660
611		661
612	Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between molecules and natural language. In <i>Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing</i> , pages 375–413.	662
613		663
614		664
615		665
616		666
617	Carl Edwards, Qingyun Wang, Lawrence Zhao, and Heng Ji. 2024. L+ m-24: Building a dataset for language+ molecules@ acl 2024. <i>CoRR</i> .	667
618		668
619		669
620	Carl Edwards, ChengXiang Zhai, and Heng Ji. 2021. Text2mol: Cross-modal molecule retrieval with natural language queries. In <i>Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing</i> , pages 595–607.	670
621		671
622		672
623		673
624		674
625	Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. <i>arXiv preprint arXiv:2402.01306</i> .	675
626		676
627		677
628		
	Yin Fang, Ningyu Zhang, Zhuo Chen, Lingbing Guo, Xiaohui Fan, and Huajun Chen. 2023. Domain-agnostic molecular generation with self-feedback. <i>arXiv preprint arXiv:2301.11259</i> .	
	Fleur M Ferguson and Nathanael S Gray. 2018. Kinase inhibitors: the road ahead. <i>Nature reviews Drug discovery</i> , 17(5):353–377.	
	Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vlad Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2024. Arcee’s mergekit: A toolkit for merging large language models. <i>arXiv preprint arXiv:2403.13257</i> .	
	Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. <i>arXiv preprint arXiv:2006.03654</i> .	
	Or Honovich, Roei Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. True: Re-evaluating factual consistency evaluation. <i>arXiv preprint arXiv:2204.04991</i> .	
	James Y Huang, Sailik Sengupta, Daniele Bonadiman, Yi-an Lai, Arshit Gupta, Nikolaos Pappas, Saab Mansour, Katrin Kirchoff, and Dan Roth. 2024. Deal: Decoding-time alignment for large language models. <i>arXiv preprint arXiv:2402.06147</i> .	
	Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hananeh Hajishirzi, and Ali Farhadi. 2022. Editing models with task arithmetic. <i>arXiv preprint arXiv:2212.04089</i> .	
	Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. Wice: Real-world entailment for claims in wikipedia. <i>arXiv preprint arXiv:2303.01432</i> .	
	Junmo Kang, Leonid Karlinsky, Hongyin Luo, Zhen Wang, Jacob Hansen, James Glass, David Cox, Rameswar Panda, Rogerio Feris, and Alan Ritter. 2024. Self-moe: Towards compositional large language models with self-specialized experts. <i>arXiv preprint arXiv:2406.12034</i> .	
	Maxim Khanov, Jirayu Burapachep, and Yixuan Li. 2024. Args: Alignment as reward-guided search. <i>arXiv preprint arXiv:2402.01694</i> .	
	Bernard Kippelen and Jean-Luc Brédas. 2009. Organic photovoltaics. <i>Energy & Environmental Science</i> , 2(3):251–261.	

678	Philippe Laban, Tobias Schnabel, Paul N Bennett, and	Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida,	729
679	Marti A Hearst. 2022. Summac: Re-visiting nli-	Carroll Wainwright, Pamela Mishkin, Chong	730
680	based models for inconsistency detection in sum-	Zhang, Sandhini Agarwal, Katarina Slama, Alex	731
681	marization. <i>Transactions of the Association for</i>	Ray, et al. 2022. Training language models to fol-	732
682	<i>Computational Linguistics</i> , 10:163–177.	low instructions with human feedback. <i>Advances in</i>	733
683	Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui	<i>neural information processing systems</i> , 35:27730–	734
684	Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei	27744.	735
685	Xiao, and Animashree Anandkumar. 2023a. Multi-	Alec Radford, Jeffrey Wu, Rewon Child, David Luan,	736
686	modal molecule structure–text model for text-based	Dario Amodei, Ilya Sutskever, et al. 2019. Lan-	737
687	retrieval and editing. <i>Nature Machine Intelligence</i> ,	guage models are unsupervised multitask learners.	738
688	5(12):1447–1457.	Rafael Rafailov, Archit Sharma, Eric Mitchell, Christo-	739
689	Zequn Liu, Wei Zhang, Yingce Xia, Lijun Wu, Sh-	pher D Manning, Stefano Ermon, and Chelsea Finn.	740
690	ufang Xie, Tao Qin, Ming Zhang, and Tie-Yan	2024. Direct preference optimization: Your lan-	741
691	Liu. 2023b. Molxpt: Wrapping molecules with	guage model is secretly a reward model. <i>Advances</i>	742
692	text for generative pre-training. <i>arXiv preprint</i>	<i>in Neural Information Processing Systems</i> , 36.	743
693	<i>arXiv:2305.10688</i> .	Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan,	744
694	Jieyu Lu and Yingkai Zhang. 2022. Unified deep	Inkit Padhi, Youssef Mroueh, and Payel Das. 2022.	745
695	learning model for multitask reaction predictions	Large-scale chemical language representations cap-	746
696	with explanation. <i>Journal of chemical information</i>	ture molecular structure and properties. <i>Nature</i>	747
697	<i>and modeling</i> , 62(6):1376–1387.	<i>Machine Intelligence</i> , 4(12):1256–1264.	748
698	Yizhen Luo, Kai Yang, Massimo Hong, Xingyi	John Schulman, Filip Wolski, Prafulla Dhariwal,	749
699	Liu, and Zaiqing Nie. 2023. Molfm: A multi-	Alec Radford, and Oleg Klimov. 2017. Proxi-	750
700	modal molecular foundation model. <i>arXiv preprint</i>	mal policy optimization algorithms. <i>arXiv preprint</i>	751
701	<i>arXiv:2307.09484</i> .	<i>arXiv:1707.06347</i> .	752
702	Timothy R McIntosh, Teo Susnjak, Tong Liu, Paul	Tal Schuster, Sihao Chen, Senaka Buthpitiya, Alex	753
703	Watters, and Malka N Halgamuge. 2024. Inadequa-	Fabrikant, and Donald Metzler. 2022. Stretch-	754
704	cies of large language model benchmarks in the era	ing sentence-pair nli models to reason over	755
705	of generative artificial intelligence. <i>arXiv preprint</i>	long documents and clusters. <i>arXiv preprint</i>	756
706	<i>arXiv:2402.09880</i> .	<i>arXiv:2204.07447</i> .	757
707	Yu Meng, Mengzhou Xia, and Danqi Chen.	Philippe Schwaller, Teodoro Laino, Théophile Gaudin,	758
708	2024. Simpo: Simple preference optimization	Peter Bolgar, Christopher A Hunter, Costas Bekas,	759
709	with a reference-free reward. <i>arXiv preprint</i>	and Alpha A Lee. 2019. Molecular transformer: a	760
710	<i>arXiv:2405.14734</i> .	model for uncertainty-calibrated chemical reaction	761
711	Mohammed Muqeeth, Haokun Liu, and Colin Raffel.	prediction. <i>ACS central science</i> , 5(9):1572–1583.	762
712	2023. Soft merging of experts with adaptive routing.	Molecular Sets. 2022. A benchmarking platform for	763
713	<i>arXiv preprint arXiv:2306.03745</i> .	molecular generation models.	764
714	Ani Nenkova, Rebecca Passonneau, and Kathleen	Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel	765
715	McKeown. 2007. The pyramid method: Incorpor-	Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford,	766
716	ating human content selection variation in summa-	Dario Amodei, and Paul F Christiano. 2020. Learn-	767
717	ri- zation evaluation. <i>ACM Transactions on Speech</i>	ing to summarize with human feedback. <i>Advances</i>	768
718	<i>and Language Processing (TSLP)</i> , 4(2):4–es.	<i>in Neural Information Processing Systems</i> , 33:3008–	769
719	Yixin Nie, Haonan Chen, and Mohit Bansal. 2019a.	3021.	770
720	Combining fact extraction and verification with neu-	Anke Tang, Li Shen, Yong Luo, Nan Yin, Lefei	771
721	ral semantic matching networks. In <i>Proceedings</i>	Zhang, and Dacheng Tao. 2024. Merging multi-	772
722	<i>of the AAAI conference on artificial intelligence</i> ,	task models via weight-ensembling mixture of ex-	773
723	volume 33, pages 6859–6866.	perts. <i>arXiv preprint arXiv:2402.00433</i> .	774
724	Yixin Nie, Adina Williams, Emily Dinan, Mo-	Kushal Tirumala, Aram Markosyan, Luke Zettle-	775
725	hit Bansal, Jason Weston, and Douwe Kiela.	moyer, and Armen Aghajanyan. 2022. Memoriza-	776
726	2019b. Adversarial nli: A new benchmark for	tion without overfitting: Analyzing the training dy-	777
727	natural language understanding. <i>arXiv preprint</i>	namics of large language models. <i>Advances in</i>	778
728	<i>arXiv:1910.14599</i> .		

779			
780		<i>Neural Information Processing Systems</i> , 35:38274–38290.	
781	Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. <i>arXiv preprint arXiv:2307.09288</i> .		
782			
783			
784			
785			
786			
787	Amos Tversky and Daniel Kahneman. 1992. Advances in prospect theory: Cumulative representation of uncertainty. <i>Journal of Risk and uncertainty</i> , 5:297–323.		
788			
789			
790			
791	Alain C Vaucher, Federico Zipoli, Joppe Geluykens, Vishnu H Nair, Philippe Schwaller, and Teodoro Laino. 2020. Automated extraction of chemical synthesis actions from experimental procedures. <i>Nature communications</i> , 11(1):3601.		
792			
793			
794			
795			
796	Ke Wang, Nikolaos Dimitriadis, Guillermo Ortiz-Jimenez, François Fleuret, and Pascal Frossard. 2024. Localizing task information for improved model merging and compression. <i>arXiv preprint arXiv:2405.07813</i> .		
797			
798			
799			
800			
801	Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In <i>Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)</i> , pages 1112–1122.		
802			
803			
804			
805			
806			
807			
808	Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In <i>International conference on machine learning</i> , pages 23965–23998. PMLR.		
809			
810			
811			
812			
813			
814			
815			
816	Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models. <i>arXiv preprint arXiv:2309.11674</i> .		
817			
818			
819			
820			
821	Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. <i>arXiv preprint arXiv:2401.08417</i> .		
822			
823			
824			
825			
826			
827	Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023. Resolving interference when merging models. <i>arXiv preprint arXiv:2306.01708</i> , 1.		
828			
829			
830			
	Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. 2024. Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities. <i>arXiv preprint arXiv:2408.07666</i> .		831 832 833 834 835
	Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. 2023. Adamerging: Adaptive model merging for multi-task learning. <i>arXiv preprint arXiv:2310.02575</i> .		836 837 838 839 840
	Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2024. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In <i>Forty-first International Conference on Machine Learning</i> .		841 842 843 844 845
	Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Dongzhan Zhou, et al. 2024. Chemllm: A chemical large language model. <i>arXiv preprint arXiv:2402.06852</i> .		846 847 848 849 850
	X Zhang, L Wang, J Helwig, Y Luo, C Fu, Y Xie, M Liu, Y Lin, Z Xu, K Yan, et al. 2023. Artificial intelligence for science in quantum, atomistic, and continuum systems. <i>arXiv 2023. arXiv preprint arXiv:2307.08423</i> .		851 852 853 854 855
	Chen Zheng, Ke Sun, Hang Wu, Chenguang Xi, and Xun Zhou. 2024. Balancing enhancement, harmlessness, and general capabilities: Enhancing conversational llms with direct rlhf. <i>arXiv preprint arXiv:2403.02513</i> .		856 857 858 859 860
	Ming Zhong, Siru Ouyang, Yizhu Jiao, Priyanka Kargupta, Leo Luo, Yanzhen Shen, Bobby Zhou, Xi-anrui Zhong, Xuan Liu, Hongxiang Li, et al. 2023. Reaction miner: An integrated system for chemical reaction extraction from textual data. In <i>Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations</i> , pages 389–402.		861 862 863 864 865 866 867 868
	Yuyan Zhou, Liang Song, Bingning Wang, and Weipeng Chen. 2024. Metagpt: Merging large language models using model exclusive task arithmetic. <i>arXiv preprint arXiv:2406.11385</i> .		869 870 871 872

873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918

A Complementary Experiments in Model Merging

For our best-performing model, CPO+SLERP, we conducted ablation studies to examine the impact of coefficients in model merging through weight interpolation of pretrained models on MoCG tasks. Specifically, we used Meditron, trained for caption-to-molecule (Cap2Mol) generation (Edwards et al., 2024), as the base model from which the merging process begins. For the source model, we deployed Meditron trained for molecule-to-caption (Mol2Cap) generation. Our experiments focused on blending weights across all layer (i.e., 0-32) from the source model into the base model while preserving Cap2Mol performance and enhancing Mol2Cap performance (see Tables 2-3), ultimately obtaining a universal model with improved overall capabilities.

We began by blending 20% of the source model’s weights with 80% of the base model’s weights, represented as $Ratio(Cap2Mol : Mol2Cap) = 1 : 4$. We then iteratively adjusted the ratio coefficient to obtain a universal model that maintained satisfactory inference performance for both tasks. Specifically, we conducted experiments with coefficient ratios of 1 : 4, 1 : 8, 1 : 16, and 1 : 32. Figure 6 overviews the experimental results.

Overall, we observed that when merging models with a relatively high percentage of weights from the source model (i.e., ratios of 1 : 4 and 1 : 8 in Figure 6), the universal model showed decreased performance on the Cap2Mol task. By contrast, when the percentage of source model weights was kept minimal (i.e., ratio of 1 : 32 in Figure 6), the universal model showed decreased performance on the Mol2Cap task. Based on these results, we concluded that the optimal ratio for merging models in MoCG tasks is 1 : 18.

We compared SLERP and TIER model merging techniques against a weighted linear combination of parameters, referred to as model soup (Wortsman et al., 2022), when applying CPO in the MoCG task. Our results indicated that model soup caused a significant drop in performance for both Mol2Cap and Cap2Mol tasks (see

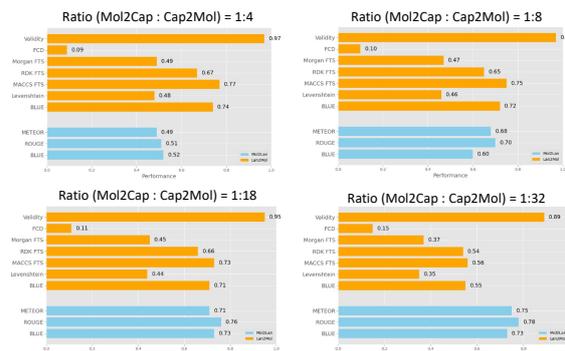


Figure 6: Inference performance for Mol2Cap and Cap2Mol tasks, achieved by merging weights from task-specific pretrained models at varying ratios to obtain universal models.

Fig. 7). We hypothesise that this is because model soup assumes that performance improvement or preservation is linearly related to weight blending, which may not hold for complex models. This observation justifies our decision to explore task-specific arithmetic and geometric merging approaches, as they inherently manage conflicts and better preserve the strengths of each model in specialised tasks.

919
920
921
922
923
924
925
926
927

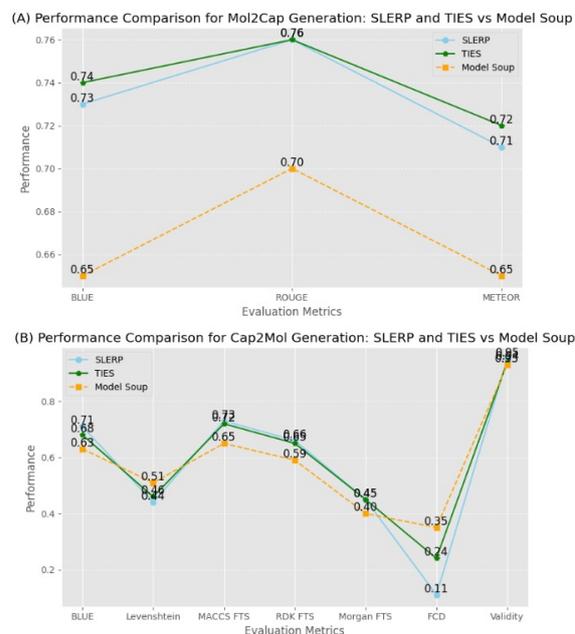


Figure 7: Comparison of SLERP and TIES with Model Soup for (A) Mol2Cap and (B) Cap2Mol generation.

928
929

B Complementary Experiments in Our Atomic-Level NLI Evaluation Method

930
931
932
933
934
935
936
937
938
939
940
941
942

We conducted ablation studies on our atomic-level NLI evaluation method to investigate potential issues in semantic understanding, such as a loss of cohesiveness in complex and lengthy captions due to excessive decomposition into atomic units. First, we analysed the distribution of word counts in captions from the test subset. We observed that the captions are typically short, with an average of 31 words (STD = 50) as shown in Fig. 8. Additionally, the captions generally exhibit little dependency across sentences, as they consist of simple natural language describing chemical properties (for a more detailed view, see Table 6).

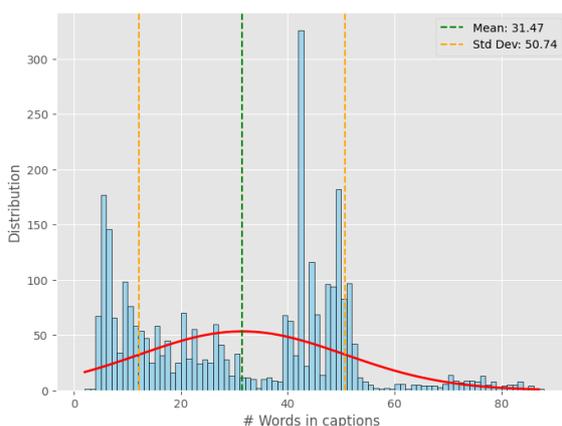


Figure 8: Distribution of word counts in captions from the test subset.

943
944
945
946
947
948
949
950
951
952
953
954
955
956
957

Based on the above word count distribution analysis, we filtered captions of varying lengths for our ablation studies: long captions (at least 50 words) and extreme captions (at least 70 words). Figures 9 and 10 illustrate the robustness of our atomic-level NLI method in comparison to other leading methods, particularly in handling long and extreme cases.

For long captions, our NLI method demonstrated a significant improvement in its ability to differentiate preferred outputs from dispreferred ones accurately, achieving a KL divergence of 2.53 (see Fig. 9), as opposed to a KL divergence of 0.54 across all cases in the test subset (see Fig. 5). In contrast, other leading NLI methods ex-

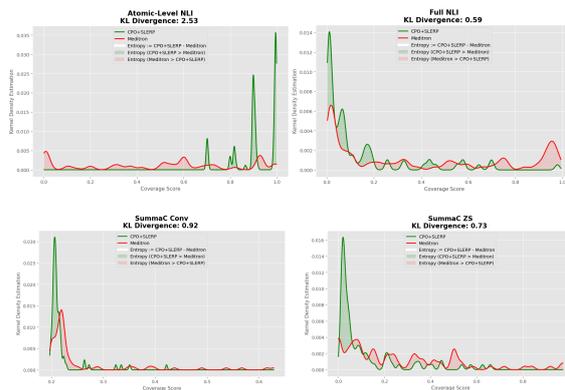


Figure 9: Relative entropy in coverage scores for preferred vs. dispreferred generated captions across atomic-level and leading NLI approaches in long captions.

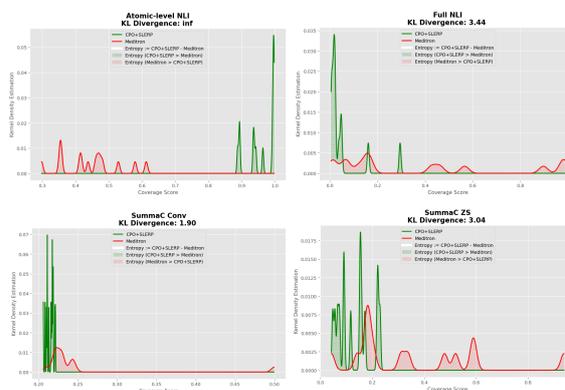


Figure 10: Relative entropy in coverage scores for preferred vs. dispreferred generated captions across atomic-level and leading NLI approaches in extreme captions.

perienced a marked increase in KL divergence, favouring dispreferred outputs, which misaligned with the entailment aspect. A similar trend was observed with extreme captions (see Fig. 10). Our ablation studies demonstrate that our atomic-level NLI method effectively handles long and complex captions in the MoCG task, whereas established NLI approaches lacked reliability in evaluating lengthy sequences.

958
959
960
961
962
963
964
965
966

C Foundations in Alignment with RLHF

Feedback-aligned LLMs traditionally undergo fine-tuning with RLHF, where human preferences serve as a reward signal in optimisation (Stiennon

967
968
969
970

et al., 2020; Ouyang et al., 2022). To train a LLM with RLHF, a reinforcement learning optimisation algorithm such as PPO (Schulman et al., 2017) is typically deployed on offline preference data, commonly involving three steps:

- **Model Training:** Typically, a model π is trained for auto-regressive language generation on a large generic corpus. This training operates under the premise that the probability distribution of a sequence of words can be broken down into the product of conditional distributions for the next word (Radford et al., 2019).
- **Reward Model Training:** A reference model π_{ref} is employed to optimise π for a downstream task. Typically, the π_{ref} model undergoes fine-tuning with an auto-regressive objective, using data pertinent to the downstream task. This often involves instruction tuning π_{ref} to regulate the generated outputs.
- **Reinforcement Learning:** The optimisation of π with respect to π_{ref} operates on a triple dataset $\mathcal{D} = \{x, y_w, y_l\}$, where x represents the input, and y_w and y_l denote preferred and dispreferred outputs, respectively, such that $y_w \succ y_l$ for x . In the Bradley–Terry model (Bradley and Terry, 1952), the probability of y_w being preferred over y_l in pairwise comparisons can be formulated as follows:

$$p^*(y_w \succ y_l | x) = \sigma(r^*(x, y_w) - r^*(x, y_l)) \quad (8)$$

Here, σ represents the logistic function, and r^* denotes the “true” reward function that underlies the preferences. As obtaining the true reward directly from a human would be prohibitively expensive, a reward model r_ϕ is trained to act as a surrogate. This is achieved by minimising the negative log-likelihood in human preference data;

$$\mathcal{L}(r_\phi) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} [\log \sigma(r_\phi(x, y_w) - r_\phi(x, y_l))] \quad (9)$$

Additionally, the Kullback-Leibler (KL) divergence between the outputs generated by π_{ref} and the parameterised π_θ models serves as an additional reward signal, ensuring that the generated responses closely align with the refer-

ence model. Consequently, an optimal model π_θ is one that maximises;

$$\mathbb{E}_{(x \in \mathcal{D}, y \in \pi_\theta)} [r_\phi(x, y)] - \beta \mathcal{D}_{\text{KL}}(\pi_\theta(y | x) || \pi_{\text{ref}}(y | x)) \quad (10)$$

where β is parameter typically $\in [0.1, 0.5]$.

Human-aware Loss Functions (HALOs):

Definition 1 (HALOs) Let $x \in X$ and $y \in Y$ denote an input and output respectively. An $f : (x, y) \rightarrow \mathbb{R}$ is considered a human-aware loss function if it satisfies

$$f(x, y; \theta) = t \left(v_f(r_\theta(x, y) - \mathbb{E}_{x' \sim Q', y' \sim Q'} [r_\theta(x', y')]) \right) \quad (11)$$

with a parameterised reward function r_θ such that $\forall (x_1, y_1), (x_2, y_2) \in X \times Y$, $r_\theta(x_1, y_1) > r_\theta(x_2, y_2) \Leftrightarrow (x_1, y_1) \succ_{r_\theta} (x_2, y_2)$, reference point distributions $Q_x(X')$ and $Q_y(Y'|X')$, a value function $v_f : \mathbb{R} \rightarrow \mathbb{R}$ that is monotonic non-decreasing and concave in $(0, \infty)$, and a negative affine function t .

RLHF can present challenges due to inherent slowness and instability, especially in the case of highly varied outputs (Zheng et al., 2024). Recently, there has been a shift towards using closed-form losses in RLHF to align LLMs with human preferences. These losses are predominantly HALOs that model human biases, as discussed in (Tversky and Kahneman, 1992), aiming to maximise the margin between preferred and dispreferred generated outputs. This approach offers a mathematical equivalence with RLHF, while effectively addressing inherent limitations.

D Data

Experiments are conducted on the benchmark *L+M-24*, which integrates molecular and linguistic modalities in four key categories: biomedical; light and electricity; human interaction and organoleptics; and agriculture and industry (Edwards et al., 2024). The dataset was created by extracting SMILES (Simplified Molecular-Input

1051 Line-Entry System) strings and their chemical
1052 properties from databases. The properties were
1053 transformed into natural language using GPT-4
1054 templates. The training and validation subsets
1055 contain around 127k and 34k language-molecule
1056 pairs, with only 10% used for our experiments.

1057 For evaluation, we randomly selected 3k un-
1058 seen pairs from a separate dataset provided by the
1059 L+M-24 research group. The samples were drawn
1060 from 635k molecule-caption pairs included in the
1061 supplementary data for the Language + Molecules
1062 @ ACL2024 workshop³.

1063 E LLM Instruction for Decomposing 1064 Text into Atomic Context Units

```
### User:  
Please breakdown the following text into indepen-  
dent facts:  
{It impacts both atherosclerosis and cardiovascular  
disease. The molecule is a nutrient, thyroxine  
treatment, fat storage that impacts pancreatitis and  
metabolic syndrome.}  
  
**  
IMPORTANT: No words or explanation is needed.  
Only include truths that are factual.  
You should NOT include any prior knowledge, and  
take the text at face value when extracting facts.  
**  
  
### Assistant:  
- The molecule impacts atherosclerosis. - The  
molecule impacts cardiovascular disease. - The  
molecule is a nutrient. - The molecule is a thyrox-  
ine treatment. - The molecule impacts pancreatitis.  
- The molecule impacts metabolic syndrome. - The  
molecule is involved in fat storage.
```

Figure 11: Instruction for decomposing text into atomic content units conveying one piece of information.

1065 F Instructions for Molecule Language 1066 Translation

³<https://github.com/language-plus-molecules/LPM-24-Dataset>

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction: You are a researcher. You can come up captions based on your existing knowledge.

Captions are given against the following input. You should be as detailed as possible.

Input: Molecule: {source molecule}
In that molecule, could you formulate a caption about?

Response: {target caption}

Instruction for caption generation, i.e., $M \rightarrow L$

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction: You are a researcher. You can come up molecule smile strings based on your existing knowledge.

Molecule smile strings are given against the following input. You should be as detailed as possible.

Input: Caption: {source caption}
In that caption, could you generate a molecule smile string?

Response: {target molecule}

Instruction for molecule generation, i.e., $L \rightarrow M$

G Baselines

- *TxtChem-T5* (Christofidellis et al., 2023) is a $T5_{XL}$ multitask model trained on linguistic and molecule modalities across multiple datasets, including CheBI-20, akin to L+M-24.
- *Chem-LLM* (Zhang et al., 2024), an InternLM2-Base-7B model, is trained on large chemical knowledge databases using DPO, achieving GPT-4-level results.
- *Meditron* (Chen et al., 2023), a 7B model, is fine-tuned on the entire L+M-24 dataset.

1078

H Evaluation Metrics

1079

1080

1081

1082

1083

1084

1085

1086

1087

1088

1089

1090

For performance evaluation, we employ established metrics from the literature (Sets, 2022; Edwards et al., 2022). In translation from molecule to language, we assess using BLEU-2, BLEU-4, ROUGE-1, ROUGE-2, ROUGE-L, and METEOR metrics. For translation from molecule to language, evaluation metrics include BLEU, Levenshtein distance, fingerprint metrics (MACCS, RDKit, and Morgan), Fréchet ChemNet Distance (FCD), and molecule validity metrics. The annotations in the result tables indicate whether higher or lower values indicate superior performance.

1091

I Training Efficiency



Figure 12: Training efficiency across alignment fine-tuning methods

1092

J Implementation Details

1093

1094

1095

1096

1097

1098

1099

1100

1101

1102

1103

1104

1105

All implementations used Meditron (Chen et al., 2023) as the backbone model, known for its performance on L+M-24. For alignment fine-tuning experiments, we initialised Meditron crossmodals, trained for molecule generation⁴. For the model merging experiments, we combined Meditron weights trained on MoCG tasks in a 1:18 ratio. This ratio aimed to preserve the balance of information between the linguistic and molecule modalities. All models were fine-tuned using QLoRA (Dettmers et al., 2024).

For the atomic-level NLI evaluation method, we instruct Meta-Llama-3-8B (Touvron et al.,

⁴Crossmodal initialisation was based on the most challenging task reported in (Edwards et al., 2024).

2023) to break down (reference, generated) pairs into a series of atomic premises and hypotheses. We then use DeBERTa⁵ to measure hallucination and coverage by performing NLI across all the atomic premises and hypotheses.

1106

1107

1108

1109

1110

```
(
  load_in_4bit=True,
  bnb_4bit_use_double_quant=True,
  bnb_4bit_quant_type=nf64,
  bnb_4bit_compute_dtype=torch.bfloat16
)
```

Figure 13: Quantisation Configurations

```
args = TrainingArguments(
  output_dir=save_path,
  overwrite_output_dir=True,
  load_best_model_at_end=True,
  num_train_epochs=3,
  per_device_train_batch_size=1
  per_device_eval_batch_size=1
  gradient_accumulation_steps=64
  gradient_checkpointing=False
  optim="adamw_torch_fused",
  learning_rate=5e-5,
  max_grad_norm=0.3,
  warmup_ratio=0.1,
  lr_scheduler_type="cosine",
)
```

Figure 14: Training configurations

```
(
  lora_alpha=16,
  r = 64,
  lora_dropout=0.1,
  task_type="CAUSAL_LM",
  bias=False,
  target_modules= "all-linear"
)
```

Figure 15: LoRA Configurations

⁵<https://huggingface.co/MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli>

1111 **K Examples of generated molecules and**
1112 **captions.**

1113 Fig. 16 and 17 illustrate examples of molecules
1114 and captions generated by our top-performing
1115 models compared to Meditron, respectively.

1116 **L Examples of Atomic-level Cross-NLI**
1117 **evaluation**

1118 Table 6 presents examples of assessing hallucina-
1119 tion and coverage in generated captions using our
1120 atomic-level cross-NLI evaluation method.

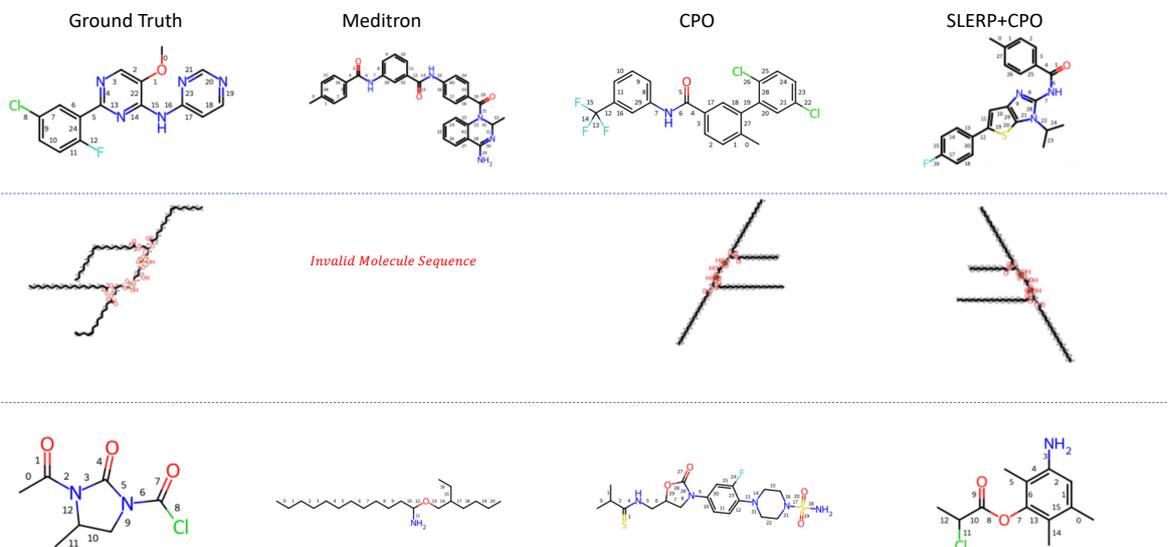


Figure 16: Examples of molecules generated by our top-performing models compared to Meditron, the best benchmark model trained on the entire dataset.

Ground Truth	Meditron	CPO	SLERP+CPO
The molecule is a stabilizing cytochrome oxidase, cholesterol translocation, proton trap for oxidative phosphorylation, apoptosis that impacts non-alcoholic fatty liver disease. The molecule is a stabilizing mitochondrial structure that impacts diabetic heart disease, aging, barth syndrome, and tangier disease	The molecule is a stabilizing cytochrome oxidase, a cholesterol translocation, and a energy source, and it impacts tangier disease. The molecule is a membrane stabilizer, a proton trap for oxidative phosphorylation, and a food additive	The molecule is a cholesterol translocation and a stabilizing cytochrome oxidase that impacts barth syndrome, non-alcoholic fatty liver disease, and aging. The molecule is a stabilizing mitochondrial structure, proton trap for oxidative phosphorylation, apoptosis that impacts tangier disease and diabetic heart disease.	The molecule is a stabilizing cytochrome oxidase and a stabilizing mitochondrial structure that impacts non-alcoholic fatty liver disease, aging, and diabetic heart disease. The molecule is a cholesterol translocation, proton trap for oxidative phosphorylation, apoptosis that impacts barth syndrome and tangier disease.
It impacts non-alcoholic fatty liver disease. The molecule is a nutrient that impacts Parkinson's disease, Alzheimer's disease, and diabetes mellitus type 2.	It impacts cardiovascular disease, Alzheimer's disease, seizure, and diabetes mellitus.	The molecule is a nutrient that impacts non-alcoholic fatty liver disease, diabetes mellitus type 2, Alzheimer's disease, and Parkinson's disease	The molecule is a nutrient that impacts non-alcoholic fatty liver disease, diabetes mellitus type 2, and Alzheimer's disease.
The molecule is a stabilizing mitochondrial structure, a stabilizing cytochrome oxidase, and a apoptosis, and it impacts diabetic heart disease. The molecule is a proton trap for oxidative phosphorylation and a cholesterol translocation, impacting both tangier disease and non-alcoholic fatty liver disease. It impacts both aging and barth syndrome.	The molecule is a stabilizing mitochondrial structure, apoptosis, and stabilizing cytochrome oxidase	The molecule is a cholesterol translocation, stabilizing cytochrome oxidase, proton trap for oxidative phosphorylation that impacts barth syndrome and non-alcoholic fatty liver disease. The molecule is a stabilizing mitochondrial structure and a apoptosis that impacts tangier disease, aging, and diabetic heart disease.	The molecule is a stabilizing cytochrome oxidase and a stabilizing mitochondrial structure that impacts non-alcoholic fatty liver disease, aging, and diabetic heart disease. The molecule is a cholesterol translocation, proton trap for oxidative phosphorylation, apoptosis that impacts barth syndrome and tangier disease.

Figure 17: Examples of captions generated by our top-performing models compared to Meditron, the best benchmark model trained on the entire dataset.

Reference Text	Atomic Premises	Generated Text	Atomic Hypothesis	Hallucination	Coverage
It impacts pancreatitis. The molecule is a fat storage and nutrient, belonging to the thyroxine treatment class of molecules, and impacts metabolic syndrome, atherosclerosis, and cardiovascular disease.	<ul style="list-style-type: none"> - The molecule impacts pancreatitis. -The molecule is a fat storage molecule. -The molecule is a nutrient. - The molecule belongs to the thyroxine treatment class of molecules. - The molecule impacts metabolic syndrome. - The molecule impacts atherosclerosis. - The molecule impacts cardiovascular disease. 	The molecule is a nutrient.	<ul style="list-style-type: none"> - The molecule is a nutrient. 	0.00	0.14
The molecule is a energy storage and is floral. The molecule is a emulsifier, nutrient, surfactant, energy source, membrane stabilizer, and rose.	<ul style="list-style-type: none"> - The molecule is a floral energy storage. - The molecule is an emulsifier. - The molecule is a nutrient. - The molecule is a surfactant. - The molecule is an energy source. - The molecule is a membrane stabilizer. - The molecule is rose. 	The molecule is a energy storage, a membrane stabilizer, and a energy source. The molecule is a surfactant, a emulsifier, and a nutrient.	<ul style="list-style-type: none"> - The molecule is an energy storage. - The molecule is a membrane stabilizer. - The molecule is an energy source. - The molecule is a surfactant. - The molecule is an emulsifier. - The molecule is a nutrient. 	0.00	0.75
The molecule is a orexin receptor antagonist.	<ul style="list-style-type: none"> - The molecule is an orexin receptor antagonist. 	The molecule is a anti viral.	<ul style="list-style-type: none"> - The molecule is an anti-viral. 	0.75	0.00
The molecule is a stabilizing cytochrome oxidase, apoptosis, stabilizing mitochondrial structure that impacts non-alcoholic fatty liver disease and tangier disease. The molecule is a cholesterol translocation and a proton trap for oxidative phosphorylation that impacts aging, barth syndrome, and diabetic heart disease.	<ul style="list-style-type: none"> - The molecule is a cytochrome oxidase. - The molecule is a stabilizer of apoptosis. - The molecule is a stabilizer of mitochondrial structure. - The molecule impacts non-alcoholic fatty liver disease. - The molecule impacts Tangier disease. - The molecule is a cholesterol translocation. - The molecule is a proton trap. - The molecule impacts oxidative phosphorylation. - The molecule impacts aging. - The molecule impacts Barth syndrome. - The molecule impacts diabetic heart disease. 	The molecule is a cholesterol translocation, a apoptosis, and a stabilizing cytochrome oxidase, and it impacts tangier disease. The molecule is a stabilizing mitochondrial structure and a proton trap for oxidative phosphorylation that impacts barth syndrome, aging, and non-alcoholic fatty liver disease. It impacts diabetic heart disease.	<ul style="list-style-type: none"> - The molecule is a cholesterol translocation. - The molecule is involved in apoptosis. - The molecule is a stabilizing cytochrome oxidase. - The molecule impacts Tangier disease. - The molecule is a stabilizing mitochondrial structure. - The molecule is a proton trap for oxidative phosphorylation. - The molecule impacts Barth syndrome. - The molecule impacts aging. - The molecule impacts non-alcoholic fatty liver disease. - The molecule impacts diabetic heart disease. 	0.00	0.91
The molecule is a anti microbial member of the anti fungal class.	<ul style="list-style-type: none"> - The molecule is anti-microbial. - The molecule is a member of the anti-fungal class. 	It belongs to the anti viral class of molecules. The molecule is both a hepatitis c treatment and a hev inhibitor.	<ul style="list-style-type: none"> - The molecule belongs to the anti-viral class of molecules. - The molecule is a hepatitis C treatment. - The molecule is an HCV inhibitor. 	0.02	0.10

Table 6: Cases showcasing insights captured by our atomic-level cross-NLI in assessing the level of hallucination and coverage in generated captions. Red highlights indicate missing information in atomic premises or invalid information in atomic hypotheses. Hallucination refers to the introduction of information absent from the reference, while coverage assesses the recall of atomic units (refer to § 3.3).