# 認知デコーディング：眼球運動による大規模言語モデルの読みやすさ制御

原田宥都 [1,2]　染谷大河 [1]　吉田遼 [1,2]　大関洋平 [1,2]
[1] 東京大学　[2] NII LLMC
{harada-yuto, oseki}@g.ecc.u-tokyo.ac.jp

## 概要

大規模言語モデルは高いベンチマーク性能や専門的で正確な回答を実現している一方で，その生成文が読者にとって必ずしも読みやすく，処理しやすいとは限らない．本研究は読解過程で生じる認知的な負荷に着目し，読み時間の推定値を読者側の処理コストの近似として用いることで，より読みやすい文章の生成を試みた．本稿では視線情報を用いた二つの介入手法を提案した．複数候補から読み時間コストの小さい出力を選び直す方法と，逐次生成中に次トークン候補の選択確率を読み時間に基づいて補正する方法である．実験では要約生成タスクを読みやすさと要約品質の指標によって評価し，さらに人間の読み時間との一致を評価した．

## 1　はじめに

近年の大規模言語モデルは知識・推論・要約など多様なタスクで高い性能を示す一方，生成文が読者にとって必ずしも処理しやすいとは限らない．とくに事後学習の過程で冗長で説明過多な文章が好まれやすいという指摘もあり [1]，正確であることに加えて，読者側の負荷を考慮した読みやすい生成文制御の重要性が増している．

生成文の読みやすさは多面的な概念であるが，本研究ではそのうち読解時の処理容易性に着目する．処理が難しい箇所では認知負荷が高まり，停滞や再読，読み飛ばしといった読解行動の変化として現れる．なかでも読み時間は理解処理の負荷と関連することが知られている [2]．そこで我々は，視線情報から得られる読み時間を認知負荷の代理指標として用いる．視線情報を NLP モデルへ取り込む試みはこれまでにも提案されており，視線や視線予測を補助信号として統合することで下流タスクを改善する研究が報告されてきた [3]．近年は視線を用いたファインチューニングにより，人間の読み時間との一致性を測る心理学的予測精度を改善できるという報告もある [4]．一方で同系統の研究では，こうした指標の改善と引き換えに，言語モデルとしての確率的整合性や下流性能，生成品質が低下するケースも報告されており，読者側の指標とモデル側の指標の間にトレードオフが生じうることが示唆されている [4]．ただし，この関係が，視線信号と尤度の本質的な競合によるものなのか，ファインチューニングに伴う分布の変化や破滅的忘却などの副作用によるものなのかは，まだ十分に整理されていない．

そこで本研究は，視線に基づく信号を用いて，モデルの重み更新を行わずに生成過程へ介入する推論時介入の枠組みを提案する．推論時介入は学習手法と比べて，介入強度を 0 にすれば元モデルへ戻れるという可逆性，ならびに介入強度を連続的に調整できるという制御性を備える．この枠組みにより，ファインチューニングに伴う副作用を回避しつつ，読者側の指標とモデル側の指標の関係をより直接に検討し，トレードオフがどの程度・どの条件で生じるかを調べる．実験では，提案手法が読みやすさや冗長性の指標を改善しつつ，人間の読み時間への一致を改善することが示された．

## 2　関連研究

### 2.1　大規模言語モデルと可読性

大規模言語モデルの生成文について，特に人間のフィードバックに基づく選好学習の文脈では，アノテーターとモデルの双方にとって長い出力や冗長な説明が好まれやすいという指摘があり [1]，これは冗長性バイアスとして議論されている [5]．この傾向は，特に医療や教育など，読み手にとっての分かりやすさが重視される領域において，生成された説明文が読み手の理解を阻害し得るという問題として

現れる [6]．可読性を扱う最も単純なアプローチは，プロンプトにより語彙の平易さや文の長さ，想定読者レベルなどのスタイル条件を与えることであり，ゼロショット指示でもある程度読みやすさの調整が可能であることが知られる [7]．しかし，この調整はユーザ側の試行錯誤に依存しやすく，学習者にとってはその操作自体が認知的な負荷となりうるほか [8]，安定した制御には反復的な生成・検証や複雑なワークフローが必要になる場合があり，いまだ容易ではない [9]．

## 2.2 視線を統合した言語処理モデル

視線は読解過程における処理負荷と関係する行動信号として長く研究されており，自然言語処理の文脈でも，視線特徴を補助情報として取り入れて文圧縮などの下流タスクを改善する試みが提案されてきた [3]．また，テキストから読み時間を推定するモデルが開発されることで，推論時であっても参照可能な信号として用いることが可能になっている [10]．さらに近年では，視線をファインチューニングや強化学習といった事後学習の枠組みに組み込む研究も現れている [4]．しかし，この学習の過程において人間の読み時間との一致性が向上する一方で，言語モデルとしての予測性能や下流タスクでの性能，生成品質が低下する可能性も報告されており [4]，読者側の指標とモデル側の指標の間にトレードオフが生じうることが示唆されている．

# 3 提案手法

本研究では，視線に由来する読み時間の推定値を読者側の処理コストの近似として用い，2 通りの方法で推論時に生成過程へ介入する．

## 3.1 Cognitive Reranking (Cog-RR)

Cognitive Reranking は，推論時に複数の候補出力を生成し，読み時間コストに基づいて最終出力を選び直す手法である．入力 $x$ に対して $N$ 個の候補列 $\{y^{(1)}, \ldots, y^{(N)}\}$ をサンプリングし，各候補について読み時間コスト $\mathrm{cost}(y^{(n)})$ を計算する．最終的に，コストが最小の候補を出力として選択する：

$$y^\star = \arg\min_{n \in \{1,\ldots,N\}} \mathrm{cost}\left(y^{(n)}\right). \tag{1}$$

コストは単純な長さ選好を避けるため，文全体の読み時間をトークン数で割った per-token の値として扱う．$N = 1$ とすれば介入しない出力と一致するため，可逆的に介入強度を調整できる．

## 3.2 Cognitive Reward-Augmented Decoding (Cog-RAD)

Cognitive Reward-Augmented Decoding は Reward-Augmented Decoding [11] に基づき，逐次生成の各時刻で次トークン候補を比較し，読み時間コストが小さい候補がサンプリングされやすくなるように logit を補正する手法である．各時刻 $t$ で top-$k$ 候補トークン集合 $C_t = \{v_1, \ldots, v_k\}$ を取り，各候補 $v_i$ について候補を付与したテキスト $x_{t,i}$ を構成し，視線モデルにより読み時間コスト $c_{t,i}$ を推定する．生成タスクでは，将来の生成が読解コストに与える影響を近似するため，$x_{t,i}$ を

$$x_{t,i} = (x, y_{<t}, v_i, \text{lookahead}) \tag{2}$$

のように prefix と候補に加えて可変長の lookahead を含む形で構成し，$c_{t,i}$ はこの先読みも含めた生成候補に対する per-token の読み時間として計算する．次に，候補集合内でコストを正規化し，読み時間が小さいほど高スコアとなるように符号を反転する．候補集合 $C_t$ における平均 $\mu_t$，標準偏差 $\sigma_t$ を用いて

$$z_{t,i} = \frac{c_{t,i} - \mu_t}{\sigma_t + \epsilon}, \qquad u_{t,i} = -z_{t,i} \tag{3}$$

と定義する．最後に，モデルの元の対数尤度 $s_t(i)$ を係数付きで補正し，

$$s'_t(i) = s_t(i) + \beta\, u_{t,i} \tag{4}$$

に基づいて次トークンをサンプリングする．ここで $\beta \geq 0$ は介入強度であり，$\beta = 1.0$ は候補集合内の標準偏差 $1\sigma$ に相当する補正を意味する．$\beta = 0$ とすれば元のデコードに一致するため，本手法も可逆的に介入強度を調整できる．

# 4 実験設定

実験では，`meta-llama/Meta-Llama-3-8B-Instruct` を用いる．以降では，評価対象となるタスク，介入手法のハイパーパラメタ，読み時間推定に用いる視線推定モデルについて述べる．

## 4.1 タスク

**要約生成（CNN/DailyMail）** 要約タスクには CNN/DailyMail を用いた．CNN/DailyMail から 500 例をランダムに抽出し，各入力記事に対して要約を生成した．生成の最大出力長は 75 トークンに制限した．プロンプトは，内容要約を指示する base prompt と，生成文の読みやすさに関するスタイル条件を追

加した readable prompt の 2 種類を用意し，比較した．これらの違いは，“Write an easy-to-read summary.” という一文を含むか否かという点のみである．評価指標として，可読性指標に Flesch Reading Ease と Flesch–Kincaid Grade Level を用いた（実装には `textstat` を用いた）．要約品質は ROUGE-1/2/L を参照要約に対して算出した．さらに，人手評価の近似として，大規模言語モデルを用いた評価である G-Eval[12] を導入した．具体的には `gpt-4o` を評価モデルとして用い，Coverage と，参照要約を与えたうえでの Conciseness をそれぞれ 1–5 の尺度で 1 回ずつ採点した．

**読み時間の予測評価（Dundee）** 人間の読み時間との一致性を測る評価のために，Dundee コーパスを用いた．テストデータから 100 例を抽出して実験した．評価設定およびプロンプト設計は栗林ら [13] を参照し，次単語予測に基づく認知モデリングの枠組みで，人間の読み時間をどの程度説明できるかを測定した．プロンプトには同論文で用いられた base の形式を採用し，“Please complete the following sentence:” という指示を用いた．

## 4.2 ハイパーパラメータ

推論時介入手法のハイパーパラメータは予備実験に基づき決定した．Cognitive Reranking（Cog-RR）では候補数を $N = 8$ とし，モデルから複数候補をサンプリングして読み時間コストが最小の候補を選択した．Cognitive Reward-Augmented Decoding（Cog-RAD）では，各時刻で介入対象とする候補集合を top-5 に固定した．また，読み時間コストの計算において，生成タスクでは候補に続く将来の生成を近似するため lookahead を導入し，要約タスクでは lookahead を 10 トークンとした．一方，次単語予測に基づく読み時間の予測評価では lookahead を 1 以上に設定するとゴールドデータのリークとなるため，0 に設定した．

## 4.3 視線推定モデル

読み時間は，人手の視線計測ではなくテキストから推定する視線推定モデルにより算出した．具体的には TorontoCL[14] と Eyettention[15] の 2 種類を用いた．TorontoCL は新聞記事データセットに付与された読み時間を学習しており，Eyettention はより汎用的な読み時間データを学習している．

本研究では，読み時間指標として First Fixation Duration (FFD) と Total Reading Time (TRT) を用いた．Eyettention はスキャンパスを出力するため，本研究ではそれを読み時間指標（FFD，TRT）へ変換し，各トークンに対する推定値として利用した．

# 5 実験結果

本節では，要約生成における読みやすさ・要約品質の変化（表 1，表 2）と，心理学的予測精度と perplexity の関係（図 1）を報告する．表中の括弧内は，同一プロンプトにおけるベースライン（介入なし）との差分である．

## 5.1 要約生成における読みやすさと要約品質

まず，プロンプトの違いだけでは読みやすさ指標は改善されなかった．今回用いたモデルでは，指示追従だけで安定して読みやすさを制御することは容易ではないことが示唆される．一方で，推論時介入を併用した場合，特に Cognitive Reranking（Cog-RR）において readable prompt で読みやすさが上がりやすい傾向が見られた．Cog-RR は複数候補の中から読み時間コストの小さい候補を選び直すため，readable prompt により「読みやすい候補」が候補集合内に出現しやすくなり，それを適切に拾えた可能性がある．これに対して Cognitive Reward-Augmented Decoding（Cog-RAD）は逐次生成中の局所的な補正であるため，プロンプトによるスタイル誘導が出力全体に一貫して反映されにくい可能性がある．要約品質については，ROUGE-1 と ROUGE-L が改善しやすい一方で，ROUGE-2 は改善が小さい，もしくは悪化する条件も見られた．この差は，介入が重要語の選択や冗長表現の抑制に寄与し，単語レベルの重なり（ROUGE-1）や最長一致部分列に基づく整合（ROUGE-L）は増えやすい一方で，連続する二語の一致（ROUGE-2）は言い換えや語順の揺れ，出力長の変化の影響を受けやすく，改善が反映されにくいことによると考えられる．特に，読み時間コストにより簡潔化が進む条件では，参照要約に含まれる重要な一単語は選べても，二語連鎖は作りづらいようである．G-Eval では，Coverage に比べて Conciseness が上がりやすい傾向が見られた．これは，読み時間コストを用いた介入が，情報の追加よりも冗長表現の削減や言い回しの圧縮に働きやすいことを示唆する．すなわち，本手法は同程度の内容をより短く表現する方向の変化を起こしたことを示唆している．

| | | | Readability | | ROUGE(↑) | | | G-Eval(↑) | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Flesch(↑) | FKGL(↓) | 1 | 2 | L | Coverage | Conciseness |
| Cog-RR | TorontoCL | FFD | 54.745$_{(+1.578)}$ | 10.345$_{(-0.837)}$ | 0.346$_{(+0.001)}$ | 0.135$_{(-0.005)}$ | 0.244$_{(-0.003)}$ | 2.900$_{(-0.025)}$ | 3.288$_{(+0.025)}$ |
| | | TRT | 53.652$_{(+0.485)}$ | 10.791$_{(-0.391)}$ | 0.350$_{(+0.006)}$ | 0.141$_{(+0.001)}$ | **0.252**$_{(+0.006)}$ | 3.000$_{(+0.075)}$ | 3.325$_{(+0.062)}$ |
| | Eyettention | FFD | 51.726$_{(-1.442)}$ | 11.631$_{(+0.449)}$ | 0.347$_{(+0.002)}$ | 0.135$_{(-0.005)}$ | 0.247$_{(+0.001)}$ | **3.075**$_{(+0.150)}$ | 3.362$_{(+0.100)}$ |
| | | TRT | **56.404**$_{(+3.236)}$ | **10.199**$_{(-0.984)}$ | 0.348$_{(+0.004)}$ | 0.137$_{(-0.003)}$ | 0.248$_{(+0.002)}$ | 3.025$_{(+0.100)}$ | 3.263$_{(+0.000)}$ |
| Cog-RAD | TorontoCL | FFD | 53.468$_{(+0.301)}$ | 11.064$_{(-0.119)}$ | 0.344$_{(-0.001)}$ | 0.138$_{(-0.003)}$ | 0.240$_{(-0.006)}$ | 2.975$_{(+0.050)}$ | 3.312$_{(+0.050)}$ |
| | | TRT | 53.518$_{(+0.350)}$ | 10.784$_{(-0.399)}$ | **0.353**$_{(+0.009)}$ | **0.145**$_{(+0.005)}$ | 0.249$_{(+0.003)}$ | 2.913$_{(-0.012)}$ | 3.312$_{(+0.050)}$ |
| | Eyettention | FFD | 53.607$_{(+0.439)}$ | 10.880$_{(-0.302)}$ | 0.342$_{(-0.003)}$ | 0.138$_{(-0.003)}$ | 0.239$_{(-0.007)}$ | 2.925$_{(+0.000)}$ | **3.350**$_{(+0.087)}$ |
| | | TRT | 53.674$_{(+0.507)}$ | 10.880$_{(-0.302)}$ | 0.342$_{(-0.003)}$ | 0.137$_{(-0.004)}$ | 0.239$_{(-0.007)}$ | 2.913$_{(-0.012)}$ | 3.337$_{(+0.075)}$ |

**表 1** Base prompt における要約結果．括弧内は同一プロンプトのベースライン（介入なし）との差分を表す．

| | | | Readability | | ROUGE(↑) | | | G-Eval(↑) | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Flesch(↑) | FKGL(↓) | 1 | 2 | L | Coverage | Conciseness |
| Cog-RR | TorontoCL | FFD | **55.985**$_{(+2.994)}$ | 10.172$_{(-0.840)}$ | 0.347$_{(+0.002)}$ | 0.139$_{(-0.001)}$ | 0.242$_{(+0.001)}$ | 2.875$_{(-0.038)}$ | 3.275$_{(+0.025)}$ |
| | | TRT | 54.325$_{(+1.334)}$ | 10.691$_{(-0.321)}$ | **0.362**$_{(+0.017)}$ | **0.146**$_{(+0.006)}$ | 0.248$_{(+0.007)}$ | **2.925**$_{(+0.012)}$ | 3.275$_{(+0.025)}$ |
| | Eyettention | FFD | 54.094$_{(+1.103)}$ | 10.720$_{(-0.292)}$ | 0.342$_{(-0.003)}$ | 0.132$_{(-0.008)}$ | 0.243$_{(+0.002)}$ | 2.888$_{(-0.025)}$ | **3.325**$_{(+0.075)}$ |
| | | TRT | 55.684$_{(+2.693)}$ | **9.950**$_{(-1.062)}$ | 0.346$_{(+0.001)}$ | 0.132$_{(-0.008)}$ | 0.239$_{(-0.002)}$ | 2.888$_{(-0.025)}$ | 3.212$_{(-0.038)}$ |
| Cog-RAD | TorontoCL | FFD | 53.716$_{(+0.725)}$ | 10.919$_{(-0.094)}$ | 0.351$_{(+0.006)}$ | 0.141$_{(+0.001)}$ | **0.250**$_{(+0.009)}$ | **2.925**$_{(+0.012)}$ | 3.275$_{(+0.025)}$ |
| | | TRT | 54.905$_{(+1.914)}$ | 10.693$_{(-0.320)}$ | 0.340$_{(-0.005)}$ | 0.138$_{(-0.001)}$ | 0.244$_{(+0.003)}$ | 2.850$_{(-0.062)}$ | 3.300$_{(+0.050)}$ |
| | Eyettention | FFD | 53.714$_{(+0.723)}$ | 10.711$_{(-0.301)}$ | 0.348$_{(+0.003)}$ | 0.139$_{(-0.001)}$ | 0.239$_{(-0.002)}$ | 2.837$_{(-0.075)}$ | 3.250$_{(+0.000)}$ |
| | | TRT | 53.645$_{(+0.654)}$ | 10.713$_{(-0.300)}$ | 0.346$_{(+0.001)}$ | 0.138$_{(-0.001)}$ | 0.239$_{(-0.003)}$ | 2.837$_{(-0.075)}$ | 3.237$_{(-0.013)}$ |

**表 2** Readable prompt における要約結果．括弧内は同一プロンプトのベースライン（介入なし）との差分を表す．

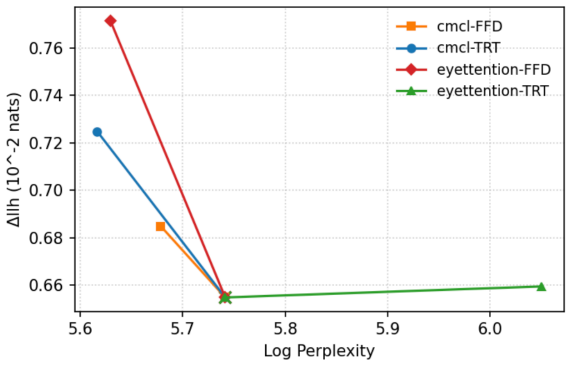


**図 1** 心理学的予測精度と perplexity の関係．ベースモデル（beta=0.0）はバツ印でプロットした．

## 5.2 心理学的予測精度と perplexity の関係

次に，人間の読み時間を説明する心理学的予測評価の結果を図 1 に示す．Cog-RAD の beta=0.0, 0.6 でのそれぞれの結果をプロットしており，beta=0.0 はつまりベースモデルと同一の結果であるため，全条件で同じ位置にプロットされる．縦軸は読み時間との一致性を測る心理学的予測精度の改善量，横軸は log perplexity である．図より，すべての条件で心理学的予測精度が改善しており，読み時間に基づく推論時介入が視線情報の説明可能性を高める方向に作用することが確認できる．さらに，そのうちの三条件では log perplexity もわずかに低下しており，読み時間の予測精度の改善が常に perplexity の悪化を伴うわけではないことが示唆される．すなわち，本設定では，読者側の指標とモデル側の指標の関係は単純なトレードオフに限られず，介入方式や読み時間推定の条件によっては両者を同時にわずかながら改善できる可能性がある．

## 6 おわりに

本稿では，大規模言語モデルの生成する文章をより読者にとって処理不可の低いものにする試みとして，視線情報を用いたモデルへの推論介入の枠組みを提案した．学習による重みの変更を介さないことにより，既存研究の報告で見られたような，読み時間に一致するほど生成文のクオリティが下がってしまうというトレードオフの現象を抑制しつつ，さらに可読性・冗長性といった指標が改善できることを示した．今後の課題として，要約以外の生成タスクへの拡張や，各指標のより大きな改善，また個人の視線情報を用いたモデルの最適化などが挙げられる．

## 謝辞

本研究は，JSPS 科研費 JP24H00087，JST さきがけ JPMJPR21C2，JST CREST JPMJCR2565，JST BOOST JPMJBY24B2 の支援を受けたものです．

## 参考文献

[1] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, **Advances in Neural Information Processing Systems**, Vol. 33, pp. 3008–3021. Curran Associates, Inc., 2020.

[2] Keith Rayner. Eye movements in reading and information processing: 20 years of research. **Psychological bulletin**, Vol. 124, No. 3, p. 372, 1998.

[3] Sigrid Klerke, Yoav Goldberg, and Anders Søgaard. Improving sentence compression by learning to predict gaze. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 1528–1533, San Diego, California, June 2016. Association for Computational Linguistics.

[4] Samuel Kiegeland, Ethan Wilcox, Afra Amini, David Robert Reich, and Ryan Cotterell. Reverse-engineering the reader. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 9367–9389, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[5] Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. Verbosity bias in preference labeling by large language models. In **NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following**, 2023.

[6] Ahmed Basharat, Rohan Shah, Nick Wilcox, Gurpaij Tur, Siddarth Tripati, Prisha Kansal, Niveah Gandhi, Sreekrishna Pokuri, Gabby Chong, Charles A Odonkor, et al. Chatgpt and low back pain-evaluating ai-driven patient education in the context of interventional pain medicine. **Interventional Pain Medicine**, Vol. 4, No. 3, p. 100636, 2025.

[7] Sean Trott and Pamela Rivière. Measuring and modifying the readability of English texts with GPT-4. In Matthew Shardlow, Horacio Saggion, Fernando Alva-Manchego, Marcos Zampieri, Kai North, Sanja Štajner, and Regina Stodden, editors, **Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)**, pp. 126–134, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[8] Maria Klar. Using chatgpt is easy, using it effectively is tough? a mixed methods study on k-12 students' perceptions, interaction patterns, and support for learning with generative ai chatbots. **Smart Learning Environments**, Vol. 12, No. 1, p. 32, 2025.

[9] Fernando Alva-Manchego, Regina Stodden, Joseph Marvin Imperial, Abdullah Barayan, Kai North, and Harish Tayyar Madabushi. Findings of the TSAR 2025 shared task on readability-controlled text simplification. In Matthew Shardlow, Fernando Alva-Manchego, Kai North, Regina Stodden, Horacio Saggion, Nouran Khallaf, and Akio Hayakawa, editors, **Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability (TSAR 2025)**, pp. 116–130, Suzhou, China, November 2025. Association for Computational Linguistics.

[10] Alberto J Molina-Cantero, Clara Lebrato-Vázquez, Juan A Castro-García, Manuel Merino-Monge, Félix Biscarri-Triviño, and José I Escudero-Fombuena. A review on visible-light eye-tracking methods based on a low-cost camera. **Journal of Ambient Intelligence and Humanized Computing**, Vol. 15, No. 4, pp. 2381–2397, 2024.

[11] Haikang Deng and Colin Raffel. Reward-augmented decoding: Efficient controlled text generation with a unidirectional reward model. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 11781–11791, Singapore, December 2023. Association for Computational Linguistics.

[12] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 2511–2522, Singapore, December 2023. Association for Computational Linguistics.

[13] Tatsuki Kuribayashi, Yohei Oseki, and Timothy Baldwin. Psychometric predictive power of large language models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, **Findings of the Association for Computational Linguistics: NAACL 2024**, pp. 1983–2005, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[14] Bai Li and Frank Rudzicz. TorontoCL at CMCL 2021 shared task: RoBERTa with multi-stage fine-tuning for eye-tracking prediction. In Emmanuele Chersoni, Nora Hollenstein, Cassandra Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus, editors, **Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics**, pp. 85–89, Online, June 2021. Association for Computational Linguistics.

[15] Shuwen Deng, David R Reich, Paul Prasse, Patrick Haller, Tobias Scheffer, and Lena A Jäger. Eyettention: An attention-based dual-sequence model for predicting human scanpaths during reading. **Proceedings of the ACM on Human-Computer Interaction**, Vol. 7, No. ETRA, pp. 1–24, 2023.