意図を踏まえた応答をする大規模言語モデルの言語間比較：6 言語を用いた検証

Cross-Linguistic Comparison of Large Language Models in Intent-Aware Responses: A Study Using Six Languages

墨 泰我 [1*]　飯田 愛結 [1]　保阪 靖人 [1]
イザベル ラヴェル [1]　小浜 正子 [1]　森山 園子 [1]　大澤 正彦 [1]
Taiga Sumi[1], Ayu Iida[1], Yasuhito Hosaka[1],
Isabelle Lavelle[1], Masako Kohama[1], Sonoko Moriyama[1], Masahiko Osawa[1]

[1] 日本大学
[1] Nihon University

Abstract: 著者らは人間の意図を踏まえた応答ができる大規模言語モデルの実現を目指し，大規模言語モデルと認知モデルの統合を提案している．しかし，使用する言語の影響で大規模言語モデルや認知モデルと統合したモデルの性能が大きく変化することがわかった．本研究では，日本語・英語・中国語・ドイツ語・イタリア語・フランス語の 6 言語を用いて意図を踏まえた応答タスクの性能を比較し，考察する．

1　はじめに

大規模言語モデル（Large Language Model: LLM）は，膨大なテキストデータを用いて訓練した深層学習モデルの一種であり，多くのコミュニケーションタスクにおいて高い性能を発揮している．本研究の目的は，大規模言語モデルを用いた自然な対話が可能な対話エージェントを，より幅広い状況で適用できるようにするために必要な知見を得ることである．

大規模言語モデルは，多様なタスクで高い性能を発揮する一方で，意図を踏まえた応答が必要なコミュニケーションタスクの性能が低いことが示されている [1]．この問題を解決するために，著者らは大規模言語モデルと認知モデル（Cognitive Model: CM）を統合する手法を提案し，その有効性を検証している [2]．

しかし，この研究では，実験に用いたプロンプトが全て日本語で書かれている点に注意する必要があった．そこで，著者らは日本語・ドイツ語・フランス語・イタリア語・中国語の 5 つの言語でプロンプトを作成し，性能を比較する実験を実施した [3]．しかしこの研究では，使用した API と実験に用いたデータセットの関係性が悪く，大規模言語モデルと認知モデルの統合手法と入力に用いる言語の関係性を調査するためには，さらなる実験が必要であった．

*連絡先：日本大学文理学部
〒 156-8550 東京都世田谷区桜上水 3-25-40
E-mail: chta22016@g.nihon-u.ac.jp

そこで本研究では，意図を踏まえた応答が必要なコミュニケーションタスクにおいて，日本語・ドイツ語・フランス語・イタリア語・中国語・英語の 6 つの言語でプロンプトを作成し，出力を言語間で比較する．実験では，先行研究 [3] で使用した大規模言語モデルから変更し，GPT-3.5-turbo を用いて，それぞれの言語で発話から意図を踏まえた応答ができるか調査した．

2　背景

2.1　大規模言語モデルと認知モデルの統合

大規模言語モデルは，意図を踏まえた応答が必要なコミュニケーションタスクにおいて，十分な性能を発揮できていないことが報告されている [1]．著者らはこの課題を解決するために，大規模言語モデルと認知モデルを統合する手法を提案し，その有効性を示している [2]．この研究では，認知モデルとして，他者の発話から他者の意図を推定し，それに基づく自己の発話を生成する，自己/他者モデル付き発話生成モデルを構築している．この認知モデルは，「自己」および「(自己が思う) 他者」それぞれの「信念 (Belief)」，「願望 (Desire)」，「意図 (Intention)」の 6 つの内部表現と「意図推定」，「意図生成」，「発話生成」の 3 つのモジュールで構成されている．

[2] では，大規模言語モデルと認知モデルを統合する

2 つの手法が提案されている．1 つ目は，認知モデル内のモジュールをそれぞれ大規模言語モデルで実装するものであり，「LLM Embedded in CM（LEC）」と名付けた手法である．2 つ目は，認知モデルを構成する内部表現やモジュールを詳細に説明するプロンプトを作成し，大規模言語モデルに入力するもので，「CM Embedded in LLM（CEL）」と名付けた手法である．

実験では，意図を踏まえた応答が必要なシチュエーションにおいて，意図を踏まえた発話を生成することができるかを検証した．具体的には，意図を踏まえた応答が必要なシチュエーションとして，「皮肉」「ツンデレ」「社会的制約」とそれぞれ名付けた 3 つのシチュエーションを作成した．さらに，4 つの条件を設定し，それらの条件間で比較実験を行った．このうち 2 つの条件は，提案手法に対応する LEC 条件および CEL 条件である．残る 2 つの条件のうち 1 つは LLM 条件であり，これは通常の大規模言語モデルに近い振る舞いをする条件である．もう 1 つは LLM with BD(LWB) 条件であり，他者の発話や自己および他者の信念・願望といった内部表現の情報は与えるが，認知モデルに関する情報は与えない条件である．

結果として，LLM 条件では全てのシチュエーションにおいて意図を踏まえた応答はなく，成功率は 0%であった．これは，設定したシチュエーションにおいて，他者の発話から意図を推定できないことを示している．内部表現を与えた LWB 条件では，ツンデレと社会的制約のシチュエーションにおいてそれぞれ 100%，90%という成功率が得られたが，皮肉シチュエーションでは 30%であった．ツンデレと社会的制約シチュエーションは，認知モデルの有効性を検証するには適さないため，皮肉シチュエーションを用いて認知モデルの有効性を検証した．皮肉シチュエーションにおいて，LEC 条件では発話生成の成功率が 100%となり，認知モデルを統合することの有効性が示された．しかし一方で，CEL 条件では発話生成の成功率は 40%にとどまった．

2.2 意図を読む大規模言語モデルの言語間の比較

大規模言語モデルは，入力する言語によって出力の傾向が左右され，特に学習データが乏しい言語では，言語の理解や生成が困難になることが示されている [4]．また，各国のコミュニケーション文化には，対話における文脈の考慮度合いに差異が存在する．文脈をより考慮する文化では，間接的かつ曖昧な表現が用いられる傾向がある一方で，文脈をあまり考慮しない文化では、直接的かつ明確な表現が用いられる特徴がある [5, 6]．このことから，第 2.1 章の実験において，プロンプトを日本語以外の言語で作成することにより，異なる傾向が示される可能性がある．

そのため，著者らはプロンプトを日本語・ドイツ語・フランス語・イタリア語・中国語の 5 つの言語で作成し，出力を比較する実験を行った [3]．プロンプトは，[2] で用いられた日本語のプロンプトを他の 4 つの言語に翻訳することで作成した．翻訳はそれぞれの言語の専門家の監訳のもとで行った．実験におけるシチュエーションや実験条件は [2] と同様のものを使用した．実験には，大規模言語モデルとして GPT-4o の API を使用した．

結果として，LLM 条件では，イタリア語の社会的制約シチュエーションで 40%となり，それ以外の言語では 0%であった．イタリア語に関しては，大袈裟に話すという文化的特徴があり，意図を読まずとも意図を踏まえたかのような発話がされた可能性があると考察している．LWB 条件では，皮肉と社会的制約シチュエーションにおいて，全ての言語で 80%以上の成功率を示したが，ツンデレシチュエーションでは，日本語・ドイツ語・フランス語・イタリア語・中国語において 80%/30%/70%/20%/50%となり，言語によって大きく異なる結果が得られた．これは，ツンデレというシチュエーション自体が日本独特の文化的側面を持つため，その馴染みやすさが文化や言語によって異なることが原因ではないかと考察している．LEC 条件と CEL 条件では，LWB 条件と比べて成功率が低いケースが見られ，特に日本語以外の言語でその傾向が顕著であった．これは，認知モデルを説明するプロンプトと言語には相性が存在するのではないかと考えられる．

この実験では，日本語の LWB 条件において，全て 80%以上の成功率となり，全てのシチュエーションにおいて大規模言語モデルと認知モデルを統合することの有効性を検証するのに適していなかった．また，この研究では英語による実験は実施しなかった．

3 実験

本実験の目的は，プロンプトに用いる言語が，意図を踏まえた応答タスクにどの程度影響を及ぼすかを検証することである．実験では，発話と発話意図に乖離のある対話を用いて，6 つの異なる言語でプロンプトを作成し，その出力を比較する．実験には，GPT-3.5-turbo を用いる．

3.1 比較言語

比較する言語として，日本語・ドイツ語・フランス語・イタリア語・中国語・英語の 6 言語を設定した．各言語の初期値および入力の例を表 1 に示す．日本語のプロンプトには，第 2.1 章で述べた [2] で用いられたも

表 1: 各言語の皮肉シチュエーションにおける初期値および入力

日本語	
他者の信念	対話相手は客である/すでに 2 時間経っている
他者の願望	早く帰ってほしい
他者の発話	「あんた、ずいぶんいい時計してはりますね〜」
自己の信念	2 時間ほどお邪魔している
自己の願望	相手に悪く思われたくない
ドイツ語	
Glaube der anderen Person	Der Gesprächspartner ist ein Kunde. / Es sind bereits 2 Stunden vergangen.
Wünsche der anderen Person	Ich möchte, dass er/sie bald weggeht.
Äußerungen der anderen Person	'Du hast ja eine ziemlich schöne Uhr da.'
Eigener Glaube	Ich bin seit etwa 2 Stunden zu Gast.
Eigene Wünsche	Ich möchte nicht, dass der Gesprächspartner schlecht von mir denkt.
フランス語	
Croyances d'autrui	L'interlocuteur est un client. / Cela fait déjà deux heures qu'il est là.
Désirs d'autrui	Je souhaite qu'il rentre rapidement.
Énoncés d'autrui	'Ah, vous avez une très belle montre !'
Croyances de soi	Je dérange depuis environ deux heures.
Désirs de soi	Je ne veux pas que l'autre pense du mal de moi.
イタリア語	
Credenze dell'Altro	L'interlocutore è un cliente / Sono già passate due ore
Desideri dell'Altro	Voglio che tu vada presto a casa
Espressioni dell'Altro	'Che bel orologio che hai!'
Credenze del Sé	Sono qui da circa due ore
Desideri del Sé	Non voglio che l'interlocutore pensi male di me
中国語	
他者的信念	对话对象是客人 / 已 经过去 2 个小 时了
他者的愿望	希望他早点回去
他者的 发言	「你这块表真不错啊」
自己的信念	已 经打 扰了大 约2 个小 时
自己的愿望	不想 让对方 对我有不好的印象
英語	
Belief of Other	The conversation partner is a customer. / Two hours have already passed.
Desire of Other	I want them to leave quickly.
Intention of Other	'You've got quite a nice watch there!'
Belief of Self	I've been staying for about two hours.
Desire of Self	I don't want the conversation partner to think badly of me.

のを利用し，他の言語のプロンプトは日本語のプロンプトを，各言語に翻訳したものを使用した．各言語への翻訳は，ドイツ語とイタリア語を専門とする第 3 著者，フランス語と英語を専門とする第 4 著者，中国語を専門とする第 5 著者の監訳のもとで行った．

3.2 実験手順

実験では，第 2.1 章で述べた [2] と同様に，4 つの条件（LLM 条件，LWB 条件，LEC 条件，CEL 条件）を設定し，各条件に対して 3 つのシチュエーション（皮肉，ツンデレ，社会的制約）での対話をそれぞれ 10 回ずつ行った．LLM 条件および LWB 条件については発話生成の出力結果を，LEC 条件および CEL 条件については意図推定・意図生成・発話生成の出力結果を，それぞれ成功/失敗/破綻と評価した．成功の基準は，他者の発話内容に基づいて意図を読み取った語句やフレーズが含まれていることである．失敗の基準は，字義通りの意味に基づいた語句やフレーズのみが含まれていることである．破綻の基準は，意図や発話が生成されていない場合，または主語が入れ替わってしまっている場合である．また，出力においてプロンプトに用いた言語以外の言語が用いられた場合も破綻とした．評価は，第 1 著者と第 2 著者が合議の上で素案を作成し，共著者に照会した．

3.3 実験結果

各シチュエーションにおける各条件下のシステムごとの成功率および破綻率を，言語ごとに表 2-7 にそれぞれ示す．

LLM 条件での成功率は，日本語のツンデレシチュエーションでのみ 20%となり，それ以外の全ての言語の全てのシチュエーションで 0%であった．また，破綻

表 2: 実験結果 日本語 成功率/破綻率 (%)

	皮肉			ツンデレ			社会的制約		
	意図推定	意図生成	発話生成	意図推定	意図生成	発話生成	意図推定	意図生成	発話生成
LLM	-	-	0/0	-	-	20/0	-	-	0/0
LWB	-	-	40/20	-	-	70/0	-	-	40/0
LEC	100/0	100/0	70/20	80/0	70/0	60/0	90/0	30/70	30/40
CEL	10/20	30/0	30/10	30/0	70/10	50/20	10/10	30/70	30/60

表 3: 実験結果 ドイツ語 成功率/破綻率 (%)

	皮肉			ツンデレ			社会的制約		
	意図推定	意図生成	発話生成	意図推定	意図生成	発話生成	意図推定	意図生成	発話生成
LLM	-	-	0/0	-	-	0/0	-	-	0/0
LWB	-	-	0/0	-	-	30/0	-	-	50/0
LEC	70/0	50/10	40/0	90/0	70/0	50/0	100/0	80/0	60/0
CEL	0/70	0/70	0/60	30/50	50/50	30/50	30/20	60/30	60/30

表 4: 実験結果 フランス語 成功率/破綻率 (%)

	皮肉			ツンデレ			社会的制約		
	意図推定	意図生成	発話生成	意図推定	意図生成	発話生成	意図推定	意図生成	発話生成
LLM	-	-	0/40	-	-	0/90	-	-	0/50
LWB	-	-	0/30	-	-	0/0	-	-	80/0
LEC	0/60	0/90	0/50	30/30	20/30	0/20	10/90	40/60	10/90
CEL	0/90	0/100	0/100	0/100	0/100	0/100	0/100	0/100	0/100

表 5: 実験結果 イタリア語 成功率/破綻率 (%)

	皮肉			ツンデレ			社会的制約		
	意図推定	意図生成	発話生成	意図推定	意図生成	発話生成	意図推定	意図生成	発話生成
LLM	-	-	0/60	-	-	0/100	-	-	0/70
LWB	-	-	10/10	-	-	40/0	-	-	0/0
LEC	10/30	30/0	0/0	70/0	0/0	0/0	90/0	30/60	30/0
CEL	0/100	0/100	0/100	0/100	0/100	0/100	0/100	0/100	0/100

表 6: 実験結果 中国語 成功率/破綻率 (%)

	皮肉			ツンデレ			社会的制約		
	意図推定	意図生成	発話生成	意図推定	意図生成	発話生成	意図推定	意図生成	発話生成
LLM	-	-	0/0	-	-	0/0	-	-	0/0
LWB	-	-	0/0	-	-	0/0	-	-	70/0
LEC	70/10	30/50	60/10	70/0	30/0	30/0	70/0	30/60	20/0
CEL	0/0	0/20	0/10	60/0	40/0	60/0	10/10	10/40	40/20

表 7: 実験結果 英語 成功率/破綻率 (%)

	皮肉			ツンデレ			社会的制約		
	意図推定	意図生成	発話生成	意図推定	意図生成	発話生成	意図推定	意図生成	発話生成
LLM	-	-	0/0	-	-	0/40	-	-	0/10
LWB	-	-	0/10	-	-	60/0	-	-	20/0
LEC	70/0	70/10	30/0	100/0	10/0	0/0	70/0	50/0	60/0
CEL	0/90	0/90	0/90	10/50	10/60	0/60	10/40	10/50	10/70

率は，日本語・ドイツ語・中国語の全てのシチュエーションで 0%となった．英語では，皮肉シチュエーションでは 0%であったが，ツンデレおよび社会的制約シチュエーションではそれぞれ 40%/10%となった．また，フランス語とイタリア語では，全てのシチュエーションで破綻率が 40%以上であった．

LWB 条件では，皮肉およびツンデレシチュエーションで日本語において最も高い成功率が得られた．一方，社会的制約シチュエーションでは，日本語よりもドイツ語・フランス語・中国語の成功率が高い結果となった．破綻率は，ツンデレおよび社会的制約シチュエーションにおいて，全ての言語で 0%となった．皮肉シチュエーションでは，日本語・フランス語・イタリア語・英語で 10% ～ 30%であった．

LEC 条件では，日本語において，皮肉およびツンデレシチュエーションにおける発話生成モジュールの成功率が，他の言語と比べて最も高いという結果が得られた．社会的制約シチュエーションでは，日本語よりもドイツ語および英語において発話生成モジュールの成功率が高くなった．破綻率は，皮肉シチュエーションにおいて，フランス語および中国語以外の全ての言語で 30%以下であった．フランス語では全てのモジュールで 50%以上，中国語では 10%/50%/10%の破綻率となったツンデレシチュエーションでは，フランス語において 20% ～ 30%の破綻が見られたが，フランス語以外の言語では破綻は見られなかった．社会的制約シ

チュエーションでは，ドイツ語および英語では全てのモジュールで 0%となった．それ以外の言語では，破綻が見られ，特に意図生成モジュールで 60% ～ 70%の破綻率となった．

CEL 条件では，皮肉シチュエーションにおいて，日本語で 10% ～ 30%の成功率が得られたが，日本語以外の全ての言語で 0%であった．また，ツンデレシチュエーションでは中国語で，社会的制約シチュエーションではドイツ語で発話生成モジュールの成功率が最も高くなった．破綻率は，フランス語およびイタリア語において全てのシチュエーションで 90%以上となった．フランス語およびイタリア語以外の言語では，皮肉シチュエーションについて，日本語および中国語で 20%以下であったが，ドイツ語および英語では 70%以上と高い破綻率となった．社会的制約シチュエーションでは 10% ～ 70%の破綻率であった．

4 考察

4.1 日本語の LEC 条件の分析

日本語の LEC 条件を 3 つのシチュエーション間で比較すると，社会的制約シチュエーションでは，意図生成モジュールと発話生成モジュールが共に 30%という低い成功率となった．社会的制約シチュエーションの出力を分析したところ，意図生成モジュールの段階で全ての出力において主語が入れ替わるという破綻が発生していた．具体的には，自己が部下で他者が上司であるシチュエーションにも関わらず，自己の意図として，「部下に対して指示を行う」といった出力がされていた．

皮肉およびツンデレシチュエーションでは，意図生成モジュールにおいて主語の入れ替わりが発生しなかったことから，社会的制約シチュエーションの初期値による影響であると考えられる．社会的制約シチュエーションの初期値では，他者の信念として「対話相手は部下」という文が設定され，意図推定モジュールに入力される．そのため，意図推定モジュールによって推定された他者の意図は，「部下に … する」というものであった．次に，推定された他者の意図は，意図生成モジュールのプロンプトの最後に追加され，大規模言語モデルに入力される．大規模言語モデルは，プロンプトの最初と最後にある文の影響を強く受けることが示されている [7]．このことから，プロンプトの最後にある「部下に … する」という文の影響を受け，主語が入れ替わってしまった可能性がある．

また，本実験とは異なるバージョンを用いた先行研究 [2, 3] において破綻が見られなかったことから，本実験で使用した GPT のバージョン（gpt-3.5-turbo）が結果に影響を及ぼした可能性も考えられる．

4.2 LWB 条件における言語間の比較

LWB 条件は，[2] で提案された認知モデルとの統合手法を用いていないため，シチュエーションと使用した大規模言語モデルのバージョンとの関係を理解する上で有用である．すなわち，心的状態を含む与えた入力から，そもそもそのバージョンのモデルが言外の意味を踏まえた応答ができるかを調べることができる．

皮肉シチュエーションでは，全ての言語の中で日本語が最も高い成功率を示した．皮肉シチュエーションは，「あんた、ずいぶんいい時計してはりますね～」という時計に言及した発話から，長い時間が経過しているため，早く帰ってほしいという意図を相手に間接的に伝えるシチュエーションとなっている．この発話は京都方言をモチーフとしているため，他の言語と比較して日本語との親和性が高かった可能性がある．

ツンデレシチュエーションでは，日本語において 70%で最も高い成功率を示した．一方，ドイツ語・フランス語・イタリア語・中国語・英語では，30% / 0% / 40% / 0% / 60%と，言語ごとに大きなばらつきが見られた．これは第 2.2 章で述べた [3] と類似する結果であった．このことから，ツンデレのように特定の国の文化的側面が強く反映されたシチュエーションでは，GPT のバージョンに関わらず，意図理解の程度が言語により大きく異なる傾向が見られる可能性がある．

社会的制約シチュエーションでは，ドイツ語・フランス語・中国語の成功率が高い結果となった．社会的制約シチュエーションは，上司が部下に対して「無理しないで」と発話するが，実際は無理してでも働いてほしいという意図を間接的に伝えるシチュエーションである．このシチュエーションは，皮肉やツンデレシチュエーションとは異なり，日本特有の文化的側面が強いシチュエーションではないため，日本語において最も高い成功率が得られなかった可能性がある．

今回の実験で使用した 3 つのシチュエーションは，日本語特有のシチュエーションが多くあるため，日本語以外の言語特有のシチュエーションでの検証も必要である．

4.3 皮肉シチュエーションにおける言語間の比較

皮肉シチュエーションでは，日本語において LWB 条件より LEC 条件の方が高い成功率を示した．また，ドイツ語・中国語・英語でも日本語と同様に LEC 条件の方が高い成功率を示した．特に中国語では，発話生成において LWB 条件では 0%という成功率となったが，LEC 条件では 60%という成功率となった．しかし，意図推定モジュールでは 70%という成功率を示したが，意図生成モジュールでは 30%という成功率を示した．出力

結果を見ると，意図生成モジュールでは破綻率が 50% であり，主語が変わった出力内容であったが，発話生成モジュールにおいて正しい主語で出力されていた．そのため，意図生成モジュールだけが低い成功率となったと考えられる．

フランス語およびイタリア語では，LEC 条件で LWB 条件より高い成功率は得られなかった．また，両言語の破綻率は，LLM 条件においてそれぞれ 40% / 60% であった．破綻率が比較的高かった要因の 1 つとして，自己の発話が出力されなかったことが挙げられる．これは，使用した GPT のバージョンを GPT-3.5-turbo に設定した影響である可能性が高い．このような破綻を回避しつつ，認知モデルを統合することの有効性を検証するためには，今回使用したモデルと，[3] で使用した GPT-4o のモデルの中間のモデルを使用することも一案として考えられる．

CEL 条件では，日本語以外の全ての言語において 0% という成功率を示した．CEL 条件は，認知モデルを詳細に 1 つのプロンプトで説明しているため，他の条件と比較してプロンプトの長さが増している．大規模言語モデルは，プロンプトのトークン数が多いほど性能が低下することが示されている [8] ため，トークン数が多い CEL 条件のプロンプトでは，LWB 条件と比べて成功率が低くなったと考えられる．大規模言語モデルの性能が向上し，長いトークン数のプロンプトに対しても性能が低下しなくなれば，CEL 条件においても性能の向上が見られるかもしれない．

5 おわりに

本研究では，大規模言語モデルが意図を踏まえた応答が必要なコミュニケーションタスクにおいて，プロンプトに用いる言語によって性能がどの程度影響を及ぼすのかを調査した．実験では，日本語・ドイツ語・フランス語・イタリア語・中国語・英語の 6 つの言語でプロンプトを作成し，その結果を比較した．今回の実験の範囲では，シチュエーションによっては LEC の手法を用いることで他の言語でも性能が向上する可能性が示唆された．

参考文献

[1] Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B. and Fedorenko, E.: Dissociating language and thought in large language models: a cognitive perspective (2023).

[2] Iida, A., Kohei, O., Satoko, F., Takashi, O., Ryoichi, N. and Masahiko, O.: Integrating Large Language Model and Mental Model of Others: Studies on Dialogue Communication Based on Implicature, in *Proceedings of the 12th International Conference on Human-Agent Interaction*, p. 260–269 (2024).

[3] 墨泰我, 飯田愛結, 長原令旺, 保阪靖人, イザベルラヴェル, 小浜正子, 森山園子, 大澤正彦：大規模言語モデルの意図理解タスクにおける多言語間の比較, 電子情報通信学会技術研究報告, 第 124 巻, pp. 34–38 (2024).

[4] Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y. and Fung, P.: A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity, in *Proceedings of the ACL*, pp. 675–718 (2023).

[5] Würtz, E.: Intercultural Communication on Web sites: A Cross-Cultural Analysis of Web sites from High-Context Cultures and Low-Context Cultures, *Journal of Computer-Mediated Communication*, Vol. 11, No. 1, pp. 274–299 (2005).

[6] Gudykunst, W. B., Matsumoto, Y., Ting-Toomey, S., Nishida, T., Kim, K. and Heyman, S.: The influence of cultural individualism-collectivism, self-construals, and individual values on communication styles across cultures, *Human Communication Research*, pp. 510–543 (1996).

[7] Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F. and Liang, P.: Lost in the Middle: How Language Models Use Long Contexts (2023).

[8] Levy, M., Jacoby, A. and Goldberg, Y.: Same Task, More Tokens: the Impact of Input Length on the Reasoning Performance of Large Language Models (2024).