第15回最先端NLP勉強会（2023.8.27）

https://aclanthology.org/2023.acl-long.524/

ACL 2023

Searching for Needles in a Haystack: On the Role of Incidental Bilingualism in PaLM's Translation Capability

Eleftheria Briakou
ebriakou@cs.umd.edu

Colin Cherry
colincherry@google.com

George Foster
fosterg@google.com

紹介者：笹野遼平（名大）

論文の概要

On the Role of Incidental Bilingualism in PaLM's Translation Capability

- Research Question:
  - 大規模言語モデル(LLMs)は意図してbilingualデータで学習していないにも関わらず高い翻訳能力(translation capability)を持つのはなぜか？
- Hypothesis:
  - 訓練事例中の意図しないバイリンガル信号(incidental bilingualism)の影響
    (44言語で3000万の翻訳ペア、ただし本仮説自体は既存[Blevins&Zettlemoyer'22])
- Contributions: PaLM (Pathways Language Model)を用いた計数・分析
  - 訓練事例に含まれるバイリンガル事例の頻度および形式 §3
  - バイリンガル事例から翻訳のための適切なpromptを発見 §4
  - バイリンガル事例が翻訳性能に与える影響 §4

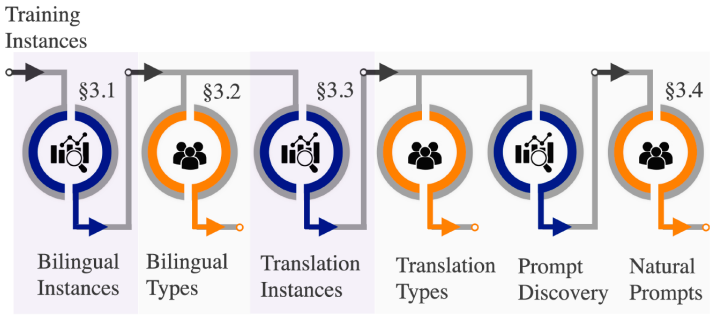
PaLM (Pathways Language Model)について

- 2022年4月にGoogleが公開した大規模言語モデル(LLM)
  - 5400億(540B)パラメータのTransformerベースの言語モデル
  - **１事例は文書中の2048までの連続トークンかそれより短い文書**
- 学習データは以下から抽出された7800億(780B)トークンで構成
  - social media conversations (50%)
  - filtered webpages (27%)
  - Wikipedia (4%)
  - books (13%)
  - News articles (1%)
  - source code (5%)

多言語混合のソース

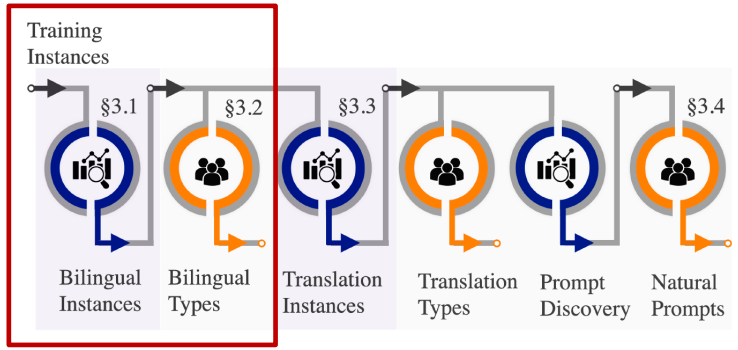
おそらく英語のみのソース

- cf. PaLM2 (2023年5月公開)
  - 英語以外の処理能力が大幅に向上、3.6Tトークン？、340Bパラメータ？



§3 Incidental Bilingualismの計数と理解

§3.1 & 3.2 バイリンガル事例の検出とタイプ分類

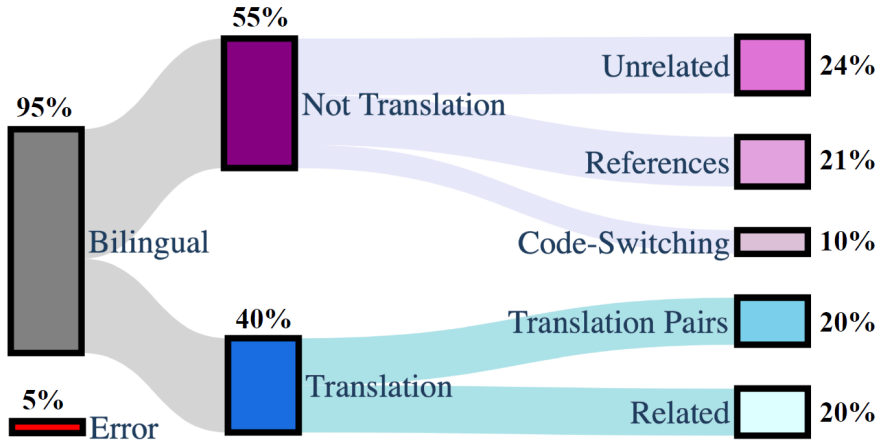


§3.1 バイリンガル事例の検出

- 言語識別モデルCMXの対象、かつ翻訳用評価データFLORES-101に含まれる44言語が対象（用例数別にHIGH:4言語, MEDIUM:11言語, LOW:29言語）
- CMXによる分析結果、以下を満たす場合にバイリンガル事例と判定
  - 10個以上の連続する英語タグ & 5個以上の連続する他言語タグが存在

§3.2 バイリンガル事例のタイプ分類

- 英仏の100事例を人手で分析
- 5%がエラー、55%が非翻訳、40%が翻訳
- 翻訳のうち半分(全体の20%)が対訳ペア、残り(20%)が含意、説明など(cf. Table 8)



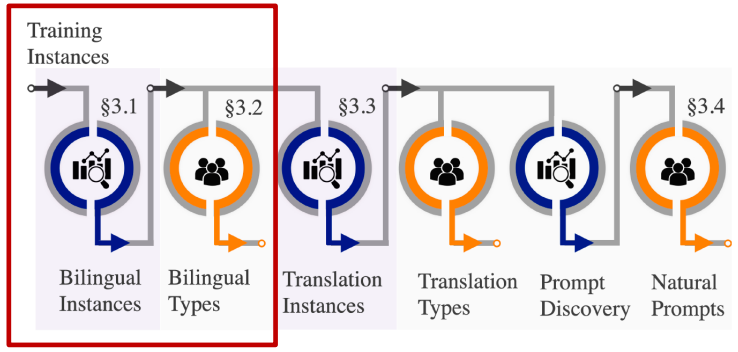
言語ごとの事例数

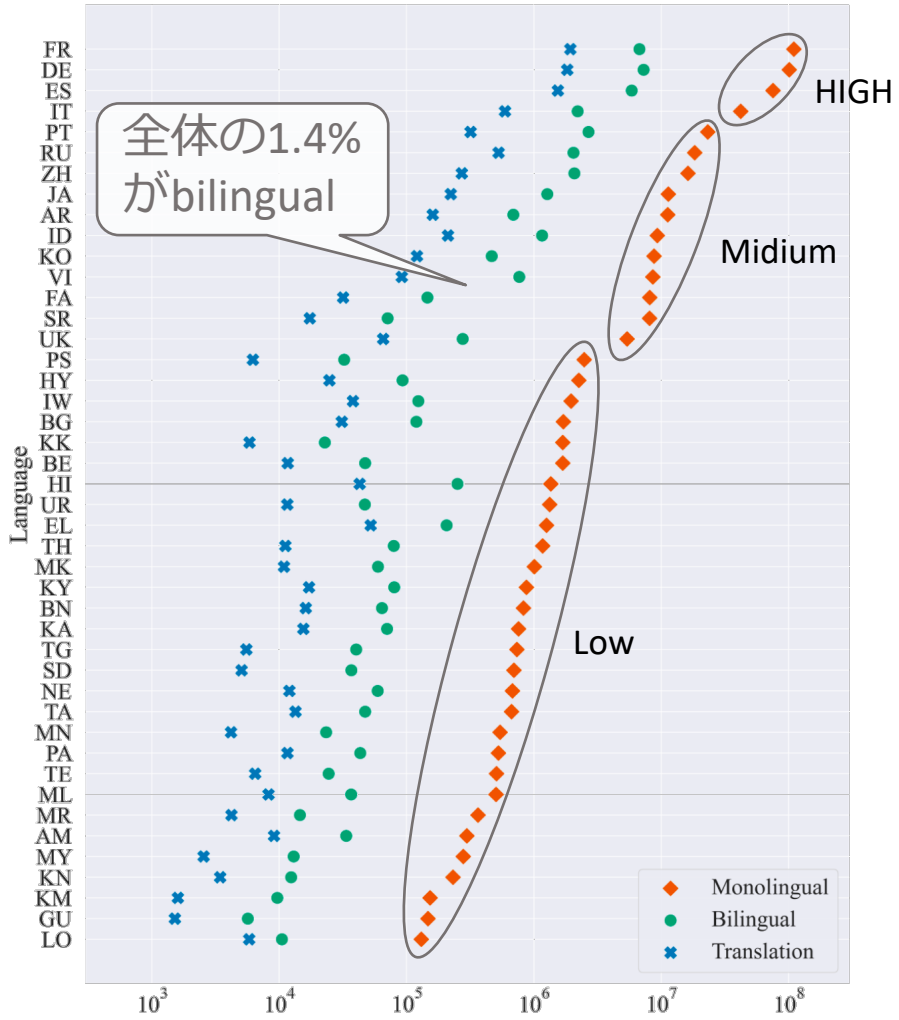
LANGUAGE	ISO	MONOLINGUAL	BILINGUAL	TRANSLATION	PARALLEL TEXTS
English	EN	2,086,622,555,000			
French	FR	109,994,921	6,743,637	1,929,032	6,618,381
German	DE	100,952,945	7,258,561	1,826,701	5,780,856
Spanish	ES	75,311,571	5,860,[illegible]34	1,538,549	5,717,352
Italian	IT	42,071,597	2,204,919	591,329	2,128,730
Portuguese	PT	23,175,895	2,685,160	317,735	1,048,717
Russian	RU	18,307,304	2,045,770	527,159	2,142,065
Chinese	ZH	16,196,482	2,075,947	271,496	706,948
Japanese	JA	11,364,144	1,271,193	222,161	601,810
Arabic	AR	11,239,689	689,215	160,554	420,851
Indonesian	ID	9,294,576	1,157,443	211,183	738,329
Korean	KO	8,777,321	465,821	120,648	248,738
Vietnamese	VI	8,588,200	767,309	91,666	265,273
Farsi	FA	8,106,752	145,498	31,685	91,731
Serbian	SR	8,092,018	70,905	17,333	49,3[illegible]
Ukrainian	UK	5,392,948	275,623	65,468	191,624
Pashto	PS	2,481,255	32,304	6,208	12,841
Armenian	HY	2,251,041	92,786	24,777	65,745
Hebrew	IW	1,956,133	123,641	37,904	111,172
Bulgarian	BG	1,702,418	119,188	30,991	83,672
Kazakh	KK	1,681,552	22,784	5,826	23,800
Belarusian	BE	1,681,272	47,284	11,646	35,535
Hindi	HI	1,356,198	250,512	42,737	121,092
Urdu	UR	1,326,867	46,973	11,564	32,654
Greek	EL	1,256,535	205,986	52,194	156,933
Thai	TH	1,169,865	79,211	11,157	28,125
Macedonian	MK	1,006,741	59,532	10,885	38,521
Kyrgyz	KY	872,384	79,955	17,107	37,484
Bengali	BN	826,933	64,012	16,138	43,046
Georgian	KA	757,142	70,220	15,457	34,939
Tajik	TG	734,888	40,146	5,503	27,889
Sindhi	SD	695,331	36,728	5,054	11,373
Nepali	NE	676,940	59,159	12,009	30,789
Tamil	TA	667,148	47,225	13,408	41,466
Mongolian	MN	541,745	23,328	4,180	12,861
Panjabi	PA	526,042	43,196	11,592	56,377
Telugu	TE	508,026	24,401	6,462	27,349
Malayalam	ML	503,762	36,652	8,235	18,412
Marathi	MR	363,791	14,544	4,209	15,684
Amharic	AM	297,463	33,604	9,098	29,355
Burmese	MY	278,933	12,989	2,547	7,020
Kannada	KN	231,308	12,386	3,430	11,589
Sinhala	KM	152,630	9,652	15,99	5,661
Gujarati	GU	146,990	5,662	1,514	5,333
Lao	LO	130,284	10,478	5,806	25,202

Table 7: Numbers of monolingual, bilingual, and translation instances across the 44 languages studied.

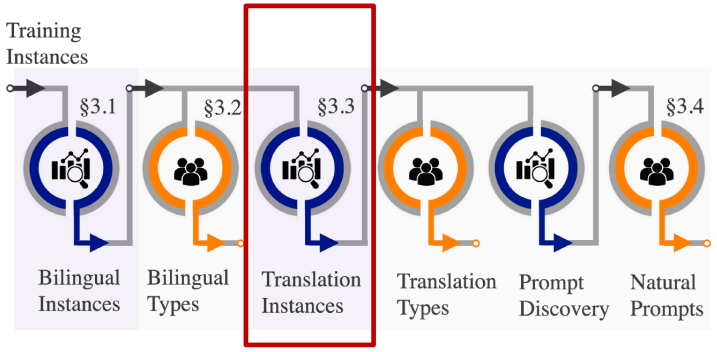
さすがに
間違いで
正しくは
×1/1000

グラフ化





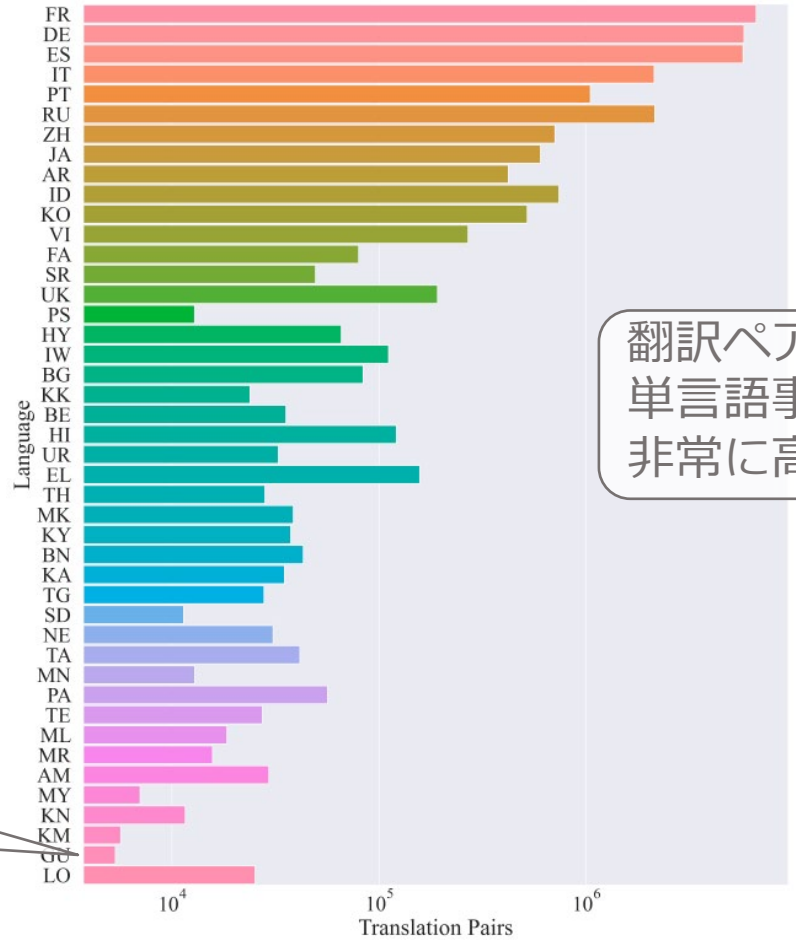
§3.3 翻訳ペアの検出



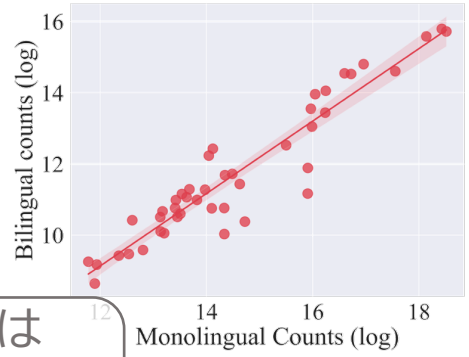
- バイリンガル事例を文に分割し、翻訳ペアを検出
- LABSEによりベクトル化し、cosine距離0.6を基準に判定

- PaLMの学習事例の0.34%が最低1つの翻訳ペアを含む

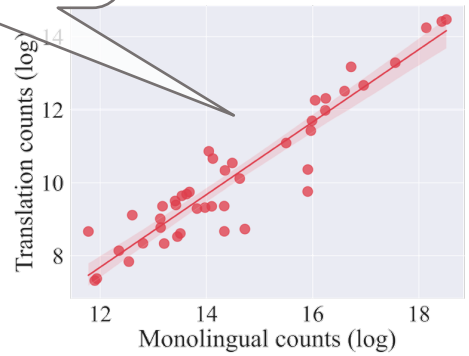
PaLMは全44言語で最低でも数千の翻訳ペアを利用



翻訳ペア数は単言語事例数と非常に高く相関

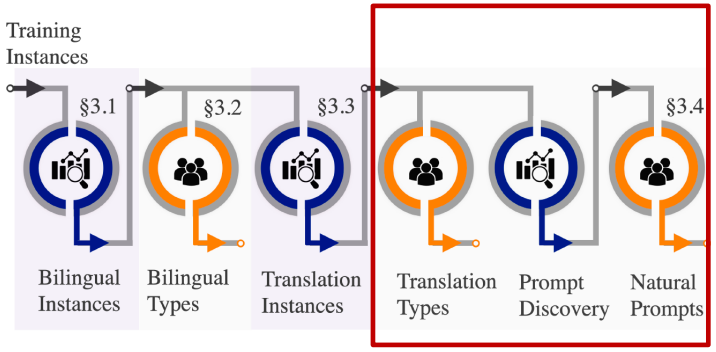


(a) $r = 0.944$



(b) $r = 0.938$

§3.4 natural promptの発見



- PaLMの学習事例中には出現するnatural (**野生の**) prompt (4タイプ)

Type	Prompt (仏語)	Prompt (日本語)
Default	French:	Japanese:
Code	FR:	JA:
Native	Français:	日本語:
Translation	Traduction:	翻訳:

	Default	Code	Native	Translation
HIGH	1,207	506	781	831
MEDIUM	219	62	136	352
LOW	38	0	64	122
ALL	1,464	568	981	1,305

Table 1: Data-driven prompt counts within PaLM's translation pairs, grouped by resourcedness.

- natural promptの分布は言語のリソース量で異なる(Table 1)
  - Codeは高リソース言語に多い
  - 低リソース言語ではNativeが多い
  - 文字種の異なる言語の扱いはやや疑問

FR	French:	Default	415
	Français:	Native	48
	Traduction:	Translation	148
	FR:	Code	177
DE	German:	Default	346
	Deutsch:	Native	407
	Übersetzung:	Translation	583
	DE:	Code	120

ZH	Chinese:	Default	60
	中文:	Native	19
JA	Japanese:	Default	21
	JA:	Code	
AR	Arabic:	Default	0

Figure 6から抜粋

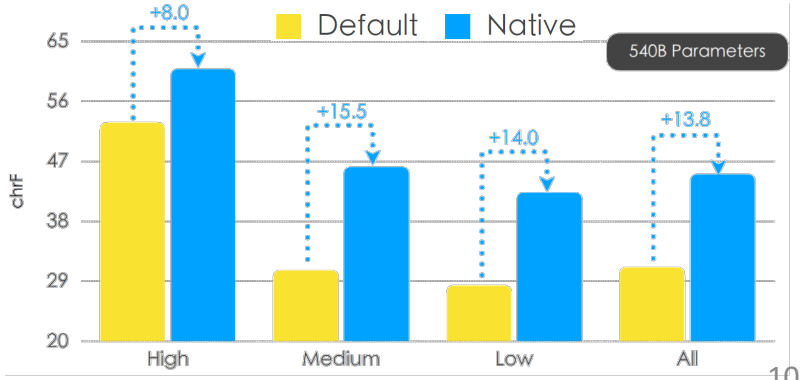
§4 Bilingualismの影響の分析

§4.1 natural promptごとの翻訳性能

- FLORES-101の評価セットを用いてPaLMの翻訳性能を検証
  - 英語への翻訳はBLEU、他言語への翻訳はchrFで評価
  - 540BパラメータのPaLMを用い、zero-shotと5-shotで実験

	Default (*zero*)		Code (*zero*)			Native (*zero*)			Translation (*zero*)			Default (*few*)		Native (*few*)		
	QUAL.	LANG.%	QUAL.	δ	LANG.%	QUAL.	δ	LANG.%	QUAL.	δ	LANG.%	QUAL.	LANG.%	QUAL.	δ	LANG.%
EN→XX																
HIGH	52.8	81.5	56.7	4.0	89.7	60.8	8.0	99.5	59.1	6.3	96.2	62.9	99.7	63.1	0.2	99.7
MEDIUM	30.6	64.8	17.2	−13.4	33.4	46.1	15.5	92.8	44.6	14.0	81.7	53.4	99.7	53.4	−0.0	99.7
LOW	28.3	69.0	2.7	−25.6	3.4	42.3	14.0	98.6	38.1	9.8	82.4	47.4	100.0	47.4	0.0	100.0
ALL	31.1	69.1	11.2	−19.9	18.8	45.0	13.8	97.2	41.6	10.5	83.5	50.3	99.9	50.3	0.0	99.9
XX→EN																
HIGH	37.6	99.7	38.5	0.9	99.6	37.7	0.1	99.7	35.4	−2.2	99.1	40.6	99.7	40.8	0.2	99.7
MEDIUM	36.9	99.5	34.8	−2.1	94.0	36.6	−0.3	99.1	35.1	−1.8	95.7	40.0	99.6	40.0	0.2	99.6
LOW	30.9	99.3	28.5	−2.3	93.7	28.4	−2.5	98.8	28.8	−2.1	90.3	35.4	99.7	35.4	0.0	99.6
ALL	33.0	99.4	31.0	−2.0	94.3	31.3	−1.7	99.0	31.0	−2.0	92.4	37.0	99.7	37.0	0.0	99.6

- 英語への翻訳ではDefaultが最良だが他も大差なし
- 他言語への翻訳ではNative、Translationが高性能、Codeは高リソース言語以外については壊滅的
- Default、Codeは生成言語を誤る傾向



EN→XX, zero-shot

§4.2 翻訳ペアの質の評価

- 翻訳ペアの量については分析したので質を評価（仏英翻訳 w/ BLEU）
  - 抽出した翻訳ペアを学習データとして、NMTモデルを0から訓練
  - LABSEベクトルのcosine距離の閾値を変えて翻訳ペアを抽出
  - モデルは6層のTransformerベースのモデル
  - WMT14の仏英タスクのデータで学習したモデルと比較
- 翻訳ペアの高い質を示す結果
  - WMTの40Mペアすべてを使った場合の41に対し、37～38のBLEUを達成
  - 訓練サイズを揃えた場合は精度差は１ポイント以内(37.3 vs 38.1)

t	#TRANSLATIONS	PaLM (mined)	WMT
N/A	40,836,876	✗	42.0
0.90	9,084,429	33.7	
0.80	7,056,441	35.7	
0.70	4,874,173	36.4	
0.60	3,341,187	37.3	38.1
0.50	2,474,703	37.2	
0.40	1,948,820	37.1	
0.30	1,477,535	38.4	36.5
0.20	906,937	37.8	
0.15	549,705	36.3	

§4.3 Incidental Bilingualismを除いた場合の翻訳性能

- Incidental Bilingualismが翻訳性能に寄与する度合いを検証
  - 翻訳事例(TRA)、バイリンガル事例(BIL)、他言語事例(NEN)を順に削除
  - 訓練データ数は同じになるように調整
  - パラメータ数1B, 8Bのモデルで実験

	ENG	NEN	BIL	TRA
FULL	43,186,985	7,224,737	517,688	270,590
-TRA	43,186,985	7,224,737	788,279	✗
-BIL	43,186,985	8,013,015	✗	✗
-NEN	51,200,000	✗	✗	✗

- 割合(≒0.6%(TRA))に比べて大きな影響
  - 何を削除したときに性能が大幅低下するかは条件によって異なる結果
  - “-NEN”でも一定の性能が出るのは言語識別エラーが原因という説明

		EN→XX (*0-shot*)				EN→XX (*5-shot*)				XX→EN (*0-shot*)				XX→EN (*5-shot*)			
		FULL	-TRA	-BIL	-NEN	FULL	-TRA	-BIL	-NEN	FULL	-TRA	-BIL	-NEN	FULL	-TRA	-BIL	-NEN
S=1B	HIGH	15.7	16.4	15.6	15.1	30.9	18.7	15.8	8.0	12.5	5.1	3.9	1.1	14.8	8.9	6.1	6.1
	MEDIUM	3.8	4.6	3.6	3.7	11.3	8.1	6.9	3.2	2.9	0.8	1.0	0.2	5.7	2.1	1.7	1.7
	LOW	0.6	0.6	0.5	0.5	6.3	6.7	5.6	3.4	0.3	0.3	0.3	0.1	0.8	0.5	0.2	0.2
	ALL	2.8	3.0	2.7	2.6	9.8	8.2	6.9	3.8	2.1	0.8	0.8	0.2	3.3	1.6	1.1	1.1
S=8B	HIGH	21.5	17.7	20.4	17.9	47.7	44.7	40.7	25.8	24.0	22.2	22.4	17.3	30.4	27.4	25.9	25.9
	MEDIUM	5.1	4.6	5.3	4.7	26.5	23.6	20.3	4.9	13.0	10.2	11.9	4.7	21.4	18.7	16.3	16.3
	LOW	1.2	0.7	1.1	0.8	8.8	8.3	7.4	2.2	2.6	2.0	2.9	0.4	6.6	5.0	4.7	4.7
	ALL	4.0	3.2	3.9	3.3	16.8	15.5	13.6	5.1	7.2	5.9	6.9	3.0	12.4	10.5	9.5	9.5

付録中の日本語関連部分を確認すると…

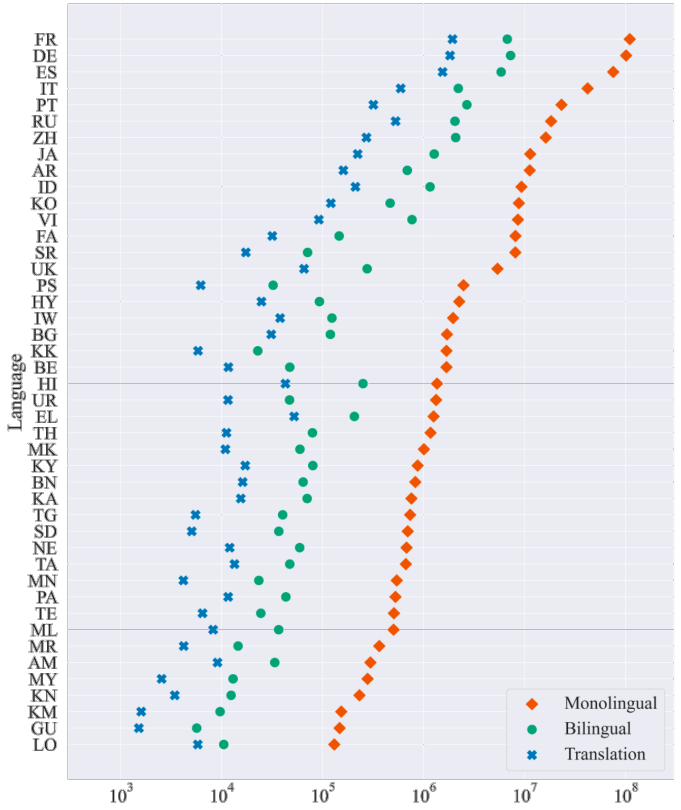
	EN→XX (*0-shot*)				EN→XX (*5-shot*)				XX→EN (*0-shot*)				XX→EN (*5-shot*)				
	FULL	-TRA	-BIL	-NEN	FULL	-TRA	-BIL	-NEN	FULL	-TRA	-BIL	-NEN	FULL	-TRA	-BIL	-NEN	Latin Script
FR	22.4	18.5	21.0	17.8	52.5	49.8	46.0	30.4	29.1	27.3	27.3	23.1	37.0	33.0	29.1	29.1	✓
DE	21.0	16.0	19.8	17.3	48.6	45.3	41.0	25.5	26.5	26.2	25.4	19.3	33.6	32.1	31.1	31.1	✓
ES	22.0	19.7	21.8	18.2	44.7	43.1	39.1	26.0	18.0	15.4	18.4	14.3	24.3	22.4	20.4	20.4	✓
IT	20.5	16.8	19.1	18.2	44.8	40.6	36.9	21.2	22.4	19.7	18.4	12.6	26.4	22.1	22.9	22.9	✓
PT	23.0	19.5	24.0	20.7	52.8	48.5	44.1	22.4	24.3	23.1	29.6	24.6	38.0	37.1	33.5	33.5	✓
RU	1.5	0.7	2.2	0.9	36.3	35.0	31.4	9.1	20.2	16.0	17.2	7.4	26.5	23.9	21.0	21.0	✗
ZH	1.6	1.4	1.5	1.3	15.7	15.3	10.4	1.0	10.9	6.5	4.9	3.5	16.2	13.6	11.2	11.2	✗
JA	1.0	0.6	1.4	0.7	12.6	10.8	8.1	1.2	6.8	6.5	4.3	1.3	13.1	9.7	7.9	7.9	✗
AR	0.8	0.8	1.4	1.2	21.7	18.2	15.6	1.8	8.5	3.6	8.4	0.6	19.9	15.9	11.7	11.7	✗
ID	15.3	15.0	14.4	15.0	45.2	41.3	37.3	8.8	19.4	16.0	18.2	9.6	28.5	23.7	22.7	22.7	✓
KO	1.7	1.8	1.8	1.4	5.4	3.7	2.9	0.4	4.7	2.6	4.1	0.8	10.5	8.2	6.1	6.1	✗
VI	8.4	8.0	8.6	7.8	34.5	30.5	23.2	3.2	9.8	6.9	9.1	1.5	19.6	18.6	13.4	13.4	✓

- ラテン文字を使う言語かは考慮
- ただし、英語と誤る可能性が考えにくい中国語(ZH)や日本語(JA)でも一定の性能
  (ZH→EN: 11.2, JA→EN: 7.9 w/ zero-shot)
- Promptの計数も一部のみ(右表)

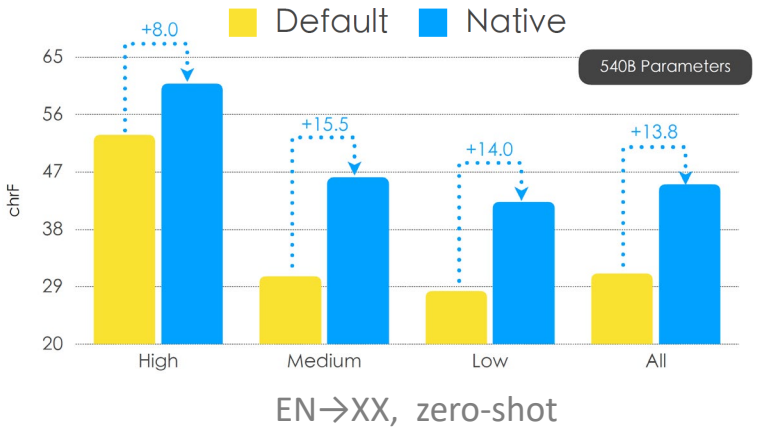
	Prompt	Type	Counts
FR	French:	Default	415
	Français:	Native	48
	Traduction:	Translation	148
	FR:	Code	177
DE	German:	Default	346
	Deutsch:	Native	407
	Übersetzung:	Translation	583
	DE:	Code	120
RU	Russian:	Default	58
	русский:	Native	20
	Перевод:	Translation	130
	RU:	Code	1
ZH	Chinese:	Default	60
	中文:	Native	19
JA	Japanese:	Default	21
	JA:	Code	

まとめ

- LLMが高い翻訳能力を持つ理由を分析
  - LLM (PaLM)を構築しているグループによる分析
  - 大量のIncidental Bilingualismが存在(44言語で3000万翻訳対)
  - natural promptが存在 & 英語への翻訳ではDefaultタイプ、英語以外への翻訳ではNativeタイプのpromptで高精度
  - 少量(<1%)の翻訳ペアが翻訳性能に大きく影響する場合も



- 気になる点
  - 公開版と比べ大幅に小規模なモデルで検証（540B vs 8B）
  - １言語は必ず英語 & 人手による分析は英仏ペアのみ
  - 文字種が異なる言語の分析の信頼性にはやや疑問



EN→XX, zero-shot