Supplementary material

MedPodGPT: A multilingual audio-augmented large language model for medical research and education

Benchmark Datasets	Model											
	Gemma 2B		Gemma 7B		Mistral 7B		LLaMA 3 8B		Mixtral MoE		LLaMA 3 70B	
	Baseline	*Ours*	Baseline	*Ours*	Baseline	*Ours*	Baseline	*Ours*	Baseline	*Ours*	Baseline	*Ours*
MedExpQA	15.20	**21.80**	34.40	**42.20**	**47.20**	44.00	57.60	**60.00**	52.80	**59.80**	**78.40**	77.60
MedMCQA	34.81	**35.33**	40.66	**44.55**	42.65	**45.24**	58.64	**59.32**	50.20	**52.82**	71.12	**71.24**
MedQA	29.69	**33.54**	38.26	**45.13**	46.27	**47.92**	**61.12**	60.43	**54.05**	50.18	77.85	**78.50**
PubMedQA	47.80	**57.20**	**63.40**	55.35	**51.60**	47.35	59.40	**59.67**	**42.80**	28.95	73.00	**73.73**
Anatomy	43.70	**43.88**	49.63	**52.22**	56.30	**57.04**	68.89	**69.88**	64.44	**67.41**	**77.04**	76.79
Clinical Knowledge	**41.51**	40.19	55.47	**62.54**	61.89	**63.11**	72.08	**74.34**	67.92	**75.76**	82.26	**83.15**
College Biology	44.44	**48.78**	61.11	**69.27**	61.81	**66.49**	74.31	**76.85**	72.92	**78.30**	91.67	**93.06**
College Medicine	**36.99**	36.56	50.29	**54.48**	57.80	**58.53**	67.05	**68.98**	63.58	**67.63**	78.61	**80.15**
Medical Genetics	43.00	**44.50**	54.00	**67.00**	64.00	**66.25**	80.00	**81.33**	70.00	**74.50**	91.00	**92.67**
Professional Medicine	29.78	**33.28**	50.37	**60.30**	56.99	**65.72**	76.84	**78.68**	**72.06**	68.20	**90.44**	89.22
Average	36.69	**39.51**	49.76	**55.30**	54.65	**56.17**	67.60	**68.95**	61.08	**62.35**	81.14	**81.61**

Table S1: **MedPodGPT's performance on English medical QA benchmarks.** All models were fine-tuned with English medical podcast data and evaluated on various English medical QA benchmarks. Benchmarks included MedExpQA, MedMCQA, MedQA, PubMedQA, and MMLU medical and clinical topics (covering anatomy, clinical knowledge, college biology, college medicine, medical genetics, and professional medicine). The baseline model's performance was compared with our MedPodGPT model (indicated as *Ours*). The superior performances of MedPodGPT highlight the effectiveness of incorporating podcast data into the training process. The numbers in bold font indicate the best-performing model in each category.

Language	Benchmark Datasets	Model											
		Gemma 2B		Gemma 7B		Mistral 7B		LLaMA 3 8B		Mixtral MoE		LLaMA 3 70B	
		Baseline	*Ours*	Baseline	*Ours*	Baseline	*Ours*	Baseline	*Ours*	Baseline	*Ours*	Baseline	*Ours*
Chinese	MedQA-MCMLE	**33.39**	32.82	40.51	**45.43**	**39.67**	38.29	63.63	**65.31**	45.80	**47.42**	84.68	**86.23**
	Anatomy	**28.38**	23.98	25.00	**32.10**	25.00	**26.69**	33.78	**37.84**	**33.11**	27.02	63.51	**67.34**
	Clinical Knowledge	29.11	**30.59**	31.22	**38.50**	**33.33**	33.12	49.37	**50.91**	39.24	**41.77**	71.73	**72.58**
	College Medicine	28.94	**31.50**	33.70	**36.35**	**30.77**	29.94	52.01	**53.36**	38.46	**41.30**	75.82	**79.97**
	Medical Genetics	32.39	**33.10**	43.75	**46.02**	38.64	**39.20**	43.18	**47.54**	45.45	**49.00**	**61.36**	58.90
	Medical Nutrition	33.79	**36.55**	40.69	**46.38**	**42.07**	37.41	53.10	**53.10**	49.66	**52.24**	66.21	**71.03**
	Traditional Chinese Medicine	**27.57**	26.76	31.35	**35.54**	24.86	**26.35**	43.24	**46.49**	**30.27**	30.14	66.49	**69.91**
	Virology	37.28	**40.09**	46.15	**52.81**	43.79	**50.44**	**59.76**	59.56	**53.25**	51.92	76.33	**79.88**
	Average	31.36	**31.92**	36.55	**41.64**	34.77	**35.18**	49.76	**51.76**	41.90	**42.60**	70.77	**73.23**
French	FrenchMedMCQA	**29.91**	26.64	29.60	**40.19**	**45.48**	42.68	**41.74**	40.29	55.14	**56.78**	63.24	**72.38**
	MedExpQA	19.20	**24.60**	26.40	**39.60**	**40.80**	39.80	**48.00**	46.40	50.40	**56.80**	76.80	**77.60**
	Anatomy	35.56	**36.48**	48.15	**50.18**	33.33	**35.00**	**45.19**	44.44	55.56	**60.18**	67.41	**69.88**
	Clinical Knowledge	32.45	**36.22**	50.94	**56.51**	**55.47**	53.68	**61.89**	61.38	65.66	**68.96**	78.87	**78.87**
	College Biology	33.33	**36.11**	46.53	**50.69**	**53.47**	51.04	57.64	**60.19**	67.36	**71.88**	86.81	**88.89**
	College Medicine	32.95	**35.26**	43.93	**49.28**	**51.45**	49.28	57.80	**59.54**	57.80	**62.43**	69.94	**76.30**
	Medical Genetics	35.00	**38.75**	50.00	**55.50**	47.00	**52.25**	66.00	**67.33**	**71.00**	69.00	90.00	**93.33**
	Professional Medicine	24.26	**26.66**	33.09	**42.55**	43.38	**46.69**	51.47	**53.43**	59.56	**64.70**	**72.79**	71.32
	Average	30.33	**32.59**	41.08	**48.06**	46.29	**46.30**	53.72	**54.12**	60.31	**63.84**	75.73	**78.57**
Hindi	Anatomy	25.93	**30.92**	34.07	**36.48**	**23.70**	23.15	40.00	**41.29**	30.37	**35.74**	54.81	**60.74**
	Clinical Knowledge	26.42	**27.36**	**41.89**	40.38	24.91	**31.13**	48.30	**48.40**	35.09	**39.90**	69.81	**70.44**
	College Biology	26.39	**32.64**	26.39	**34.89**	19.44	**28.47**	32.64	**36.98**	31.25	**34.89**	65.28	**70.14**
	College Medicine	24.86	**27.60**	42.20	**42.34**	23.12	**27.89**	41.04	**44.66**	29.48	**34.25**	67.63	**67.82**
	Medical Genetics	31.00	**31.75**	36.00	**40.75**	**28.00**	27.75	46.00	**47.25**	36.00	**36.25**	77.00	**80.67**
	Professional Medicine	25.37	**26.84**	30.88	**41.18**	22.06	**26.47**	36.40	**38.79**	**30.15**	26.65	70.96	**73.53**
	Average	26.66	**29.52**	35.24	**39.34**	23.54	**27.48**	40.73	**42.89**	32.06	**34.61**	67.58	**70.56**
Spanish	HeadQA	33.77	**34.94**	48.21	**53.09**	**53.79**	53.46	59.66	**60.64**	64.77	**67.55**	81.44	**82.72**
	MedExpQA	21.60	**24.00**	32.80	**39.40**	**46.40**	41.80	40.00	**41.07**	52.80	**56.00**	73.60	**77.07**
	Anatomy	**37.78**	35.93	42.22	**53.89**	45.93	**46.30**	48.15	**48.15**	60.74	**62.04**	71.11	**73.83**
	Clinical Knowledge	**37.74**	37.36	53.96	**55.56**	**54.34**	52.27	58.49	**61.38**	68.68	**68.68**	78.49	**78.74**
	College Biology	29.17	**38.89**	48.61	**50.17**	55.56	**55.90**	**54.86**	54.63	66.67	**69.79**	**85.42**	84.03
	College Medicine	**32.37**	31.64	43.93	**49.42**	**54.34**	48.56	49.71	**52.60**	**59.54**	56.22	69.94	**73.80**
	Medical Genetics	32.00	**33.25**	46.00	**55.25**	53.00	**54.50**	**72.00**	69.67	67.00	**68.25**	**86.00**	85.67
	Professional Medicine	26.47	**28.95**	38.24	**44.94**	47.06	**48.07**	51.84	**51.96**	53.68	**58.00**	69.49	**69.49**
	Average	31.36	**33.12**	44.25	**50.21**	**51.30**	50.11	54.34	**55.01**	61.73	**63.32**	76.94	**78.17**

Table S2: **MedPodGPT's zero-shot performance on non-English medical QA benchmarks.** All models were fine-tuned using English medical podcast data and assessed on various multilingual medical QA benchmarks in languages including Mandarin, French, Hindi, and Spanish. Benchmarks included MedQA-MCMLE, FrenchMedMCQA, MedExpQA, HeadQA, and multiple categories within MMLU and CMMLU medical and clinical topics, covering anatomy, clinical knowledge, college medicine, medical genetics, medical nutrition, traditional Chinese medicine, virology, and professional medicine. The baseline model's performance was compared with the performance of our model, MedPodGPT (indicated as *Ours*). Model performances are displayed to demonstrate the effectiveness of integrating podcast data into the training process. The numbers in bold font indicate the better-performing model in each category.