Supplementary Information

Table of contents

1 Data

1.1 Original data sample

- 医馆: 珠江馆
- 性别: 女
- 年龄: 12 岁
- 病历号: 44830
- 问: 三诊诊疗经过：自幼汗多如水，动则尤甚，头部明显，头发可滤出水来，大便成型，夏天出汗尤为明显。今日来就诊，步行 1000 步，已经大汗淋漓，衣服、裤子打湿。总体出汗较前减少 4 成。
- 诊断：1-中医诊断：自汗病
- Rx: 处方 1：大约总重量 2100.0g 炮附子 10.00g, 肉桂 10.00g, 熟地黄 60.00g 炒山药 30.00g, 山茱萸 30.00g, 牡丹皮 15.00g 茯苓 30.00g, 泽泻 30.00g, 黄芪 45.00g 焦山楂 10.00g, 焦神曲 10.00g, 焦麦芽 10.00g 当归头 10.00g// 共 7 剂（柒剂）用法：加冷水 1200ml，文火煮取 300ml，分 2 次早晚饭后温服。
- 治疗: nan
- 医生: 3
- 就诊日期: 2022-06-18
- 病案: nan
- 医嘱: nan
- 编号: 15626.0
- 望: 舌淡苔白稍腻
- 闻: nan
- 切: 脉象: 中取滑

1.2 Final data sample

- 医馆: 珠江馆
- 性别: 女
- 年龄: 12 岁
- 病历号: 44830
- 望: 舌淡苔白稍腻
- 闻: nan
- 问: 三诊诊疗经过：自幼汗多如水，动则尤甚，头部明显，头发可滤出水来，大便成型，夏天出汗尤为明显。今日来就诊，步行 1000 步，已经大汗淋漓，衣服、裤子打湿。总体出汗较前减少 4 成。
- 切: 脉象为中取滑
- 中医诊断: '自汗病'
- 基础方: 地黄丸（六味地黄丸）: 泽泻, 熟地黄, 牡丹皮, 茯苓, 山药, 山萸肉
- 医生: 3
- 就诊日期: 2022-06-18
- 编号: 15626.0
- 处方: 附子, 肉桂, 熟地黄, 山药, 山萸肉, 牡丹皮, 茯苓, 泽泻, 黄芪, 山楂, 六神曲, 麦芽, 当归

1.3 Demographics

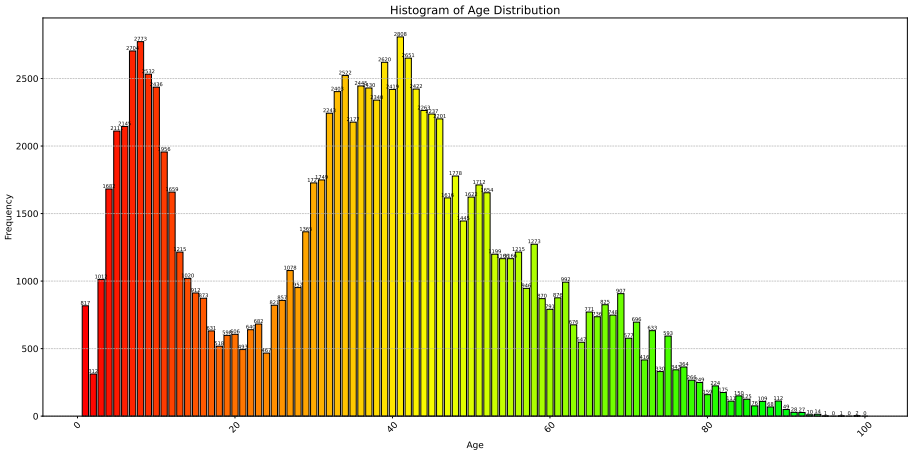

Fig. 1: Age distribution of data.

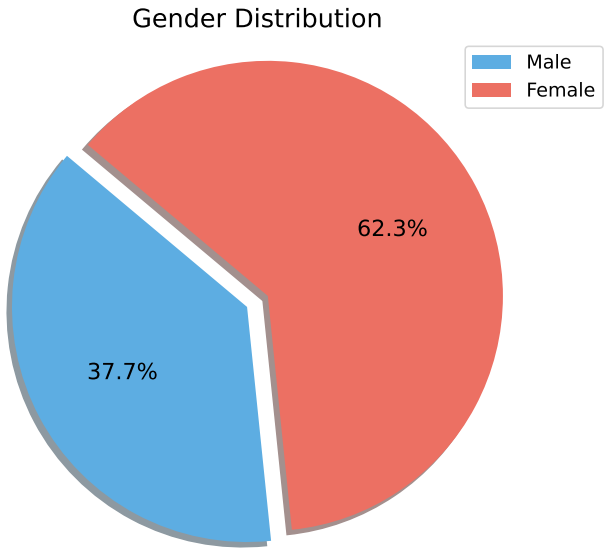

Fig. 2: Gender distribution of data.

1.4 Examples of each task

- Task 1.
  - instruction: 请你根据患者情况，给出中医的诊断结果。
  - input: 这是一条缺失了部分数据的医院的诊疗记录。患者年龄 44 岁，性别为女。由 id 为 0 的中医医生接诊。中医望诊的结果显示：舌淡暗红，胖大，苔根薄浊腻，舌络稍显粗长，有 8 个甲印，甲床红润。中医问诊的结果显示：四诊诊疗经过：近 2 月咽喉有异物感，按咽炎、胃-食管返流治疗，现无反酸、胃胀，但仍有咽喉异物感。10 余年前开始尿频、尿不尽，近 2 月加重，有漏尿，下午至晚上小便每小时 1 次。夜间陪伴小孩睡觉，睡眠常被打断。恐惧症病史。11 月 18 日胃肠镜提示：慢性萎缩性胃窦炎，直肠粘膜慢性炎症性改变。刻诊：咽喉异物感减轻，觉得咽中有痰，口干，尿频，近 3 年经常流泪，行多次泪管通畅术无明显缓解，觉疲倦，身体沉重。中医切诊的结果显示：脉象为寸关中取略滑，尺脉沉细。
  - output: 梅核气病
  - system: 你是一个经验丰富的老中医，同时也精通西医。
- Task 2.
  - instruction: 请你根据患者情况和诊断结果，给出一个对应的基础方。
  - input: 医生编号为 0。患者年龄 82 岁，性别为男。中医望诊的结果是：舌印：中，腮印：6；甲印：无。糖尿病、高血压病病史十余年以上，现血压、血糖控制平稳。形体偏胖，言语稍欠流利。中医闻诊的结果是：-。中医问诊的结果是：主诉：双下肢无力、行走困难 1 月余，伴有尿频、尿急、尿不尽、遗尿。诊疗经过：近一个多月来双下肢无力、行走困难，伴有尿频、尿急、尿不尽、遗尿，夜尿频数，甚则尿失禁，无口干口渴，困倦欲寐，晨起头晕，站立不稳，无天旋地转感。大便正常，出汗不多。2011 年出现脑梗塞，2014 年心脏植入支架 1 枚。脉象：弦，寸弱。中医诊断：痿病。
  - output: 1. 小半夏加茯苓汤：茯苓、半夏、生姜；
  - system: 你是一个经验丰富的老中医，同时也精通西医。

- Task 3.
  - instruction: 请你根据患者情况和诊断结果和基础方，给出一个对应的处方
  - input: 医生编号为 0。患者年龄 86 岁，性别为男。中医望诊的结果是：舌印：无，腮印：无，甲印：2；舌淡暗红胖大满口，苔中心白浊腻，舌面有细裂纹，舌络细如树枝状，甲床灰白。中医问诊的结果是：主诉：检查提示血三系减少 1 天。诊疗经过：2013 年 5 月曾因头晕在广州市第一人民医院住医治疗，诊断为 1. 多发腔隙性脑梗塞；2. 冠心病陈旧心肌梗死（1）PCI 术后，（2）心功能 级；3. 型糖尿病；4. 右股骨颈骨折术后；5. 肝硬化失代偿期伴脾大，血三系减少；6. 周围动脉硬化闭塞。糖尿病史 30 年，近十年注射胰岛素控制血糖，目前血糖控制尚可。今日查血常规提示：白细胞 2.86*10E9/L，红细胞 3.34*10E12/L，血红蛋白 68.2g/L，血小板 107*10E9/L。易腹泻，偶有大便失禁，小便通畅，偶有胸。脉象：左沉弦略劲右沉细弦。中医诊断：肝痹病；胸痹心痛病；消渴病。对应的基础方为（按编号排列）：1. 四苓散：猪苓、白术、茯苓、泽泻、桂枝；
  - output: 处方为（按编号排列）：1.：黄芪、当归、鸡血藤、白术、山药、炮姜、茯苓、猪苓、泽泻、桂枝、甘草、红参、鳖甲、牡蛎、阿胶、鹿角胶、龟甲胶；
  - system: 你是一个经验丰富的老中医，同时也精通西医
- Task 4.
  - instruction: 请你根据患者的处方结果，反推出对应的中医诊断结果
  - input: 医生编号为 0。患者年龄 92 岁，性别为男。中医望诊的结果是：舌印：无，腮印：无，甲印：无，舌暗红苔根稍腻（照片）。中医问诊的结果是：四诊诊疗经过（邮箱问诊）：一般活动尚可，但易疲劳，步行 2000 步便感到劳累，急走、活动量大或上 3 层楼即气促、心慌、胸闷。近几年容易感冒，穿脱不及时，不慎伤风，偶感凉、热，轻则鼻塞、流涕、打喷嚏、头痛、咳嗽，重则肺炎。平素易发口腔溃疡、口疮，牙龈肿痛。腰腿疲软、酸痛，手脚冰

凉。饱食则胃胀、反酸，大便秘结，小便时有灼热、胀痛。夜间醒来时有微汗。脉象：左沉缓，尺无力（中医师代诊）右和缓。对应的基础方为（按编号排列）：1. 培元固本散：鹿茸、琥珀、五灵脂、紫河车、三七、人参；处方为（按编号排列）：1.：紫河车、人参、三七、琥珀、鹿茸、丹参、水蛭、大黄、贝母、刺猬皮、穿山甲、五灵脂、麦芽、肉桂；

  - output: 胸痹心痛病。
  - system: 你是一个经验丰富的老中医，同时也精通西医

- Task 5.

  - instruction: 请你根据患者情况和诊断结果，补全对应的处方
  - input: 医生编号为 0。患者年龄 9 岁。性别为男。中医望诊的结果是：舌印：无，腮印：无，甲印：6；小，甲床红润，舌红，苔薄白，根稍腻。中医问诊的结果是：预诊：身高 110cm；体重 22.13Kg；过敏史：否认。既往史：否认。家族史：否认。二诊：上气、清嗓次数减少，大便 1 周 2 行，不硬，小便黄，晨起喷嚏明显，睡时有鼾声。咽后壁滤泡较多。脉象：中取滑。中医诊断：鼾症。对应的基础方为（按编号排列）：1. 小半夏加茯苓汤：生姜、半夏、茯苓；处方为（按编号排列）：1.：党参、_____、茯苓、半夏、生姜、厚朴、紫苏梗、石菖蒲、六神曲、细辛、海螵蛸、白芷；
  - output: 白术；
  - system: 你是一个经验丰富的老中医，同时也精通西医

- Task 6.

  - instruction: 请你根据患者的处方结果，反推出患者对应的年龄和性别
  - input: 医生编号为 0。中医望诊的结果是：舌印：无，腮印：无，甲印：无，舌红苔薄白腻。中医问诊的结果是：预诊：过敏史：否认。既往史：否认。家族史：否认。主诉：诊疗经过：咳嗽，有痰，咽痒，昨日喝温酒后，咽痒及咳嗽好转，但声音仍沙哑；大便通畅；脉象：弦无力。中医诊断：失音病。对应的基础方为（按编号排列）：1. 桔梗汤：桔梗、甘草；处方为（按编号排列）：1.：木蝴蝶、竹蜂、桔梗、甘草；

– output: 女,40 岁。

– system: 你是一个经验丰富的老中医，同时也精通西医

1.5 Supplementary Explanation of Data Processing

Data governance for HGipt includes terminology harmonization (herb names, syndrome terms) to reduce stylistic differences across doctors; records with diagnosis–prescription mismatch or numeric anomalies are reviewed by a senior TCM physician and then corrected or removed; deduplication uses doctor's id, patient's id and four-examinations content as the key, retaining the most complete record.

2 Results

We will use lr (learning rate) in tables.

2.1 Basic formulas mentioned at the main passage

Basic formula name	**Code nme**	**Composition**
七气汤 (Decoction of the Seven Qi)	BF0	[‘半夏’, ‘人参’, ‘生姜’, ‘桂枝’, ‘甘草’]
三拗汤 (Three Obstruction Decoction)	BF1	[‘甘草’, ‘麻黄’, ‘苦杏仁’, ‘生姜’]
不寐病指南 (Insomnia Treatment Guidelines)	BF2	[‘柴胡’, ‘赤芍’, ‘枳壳’, ‘甘草’, ‘茯神’, ‘丹参’]
丹参饮 (Danshen Drink)	BF3	[‘丹参’, ‘檀香’, ‘砂仁’]
举卿古拜散 (Juqing Gubai Powder)	BF4	[‘荆芥’]
二妙散 (Two Marvels Powder)	BF5	[‘黄柏’, ‘苍术’, ‘生姜’]
人参败毒散（败毒散）(Ginseng Detoxification Powder (Detoxification Powder))	BF6	[‘柴胡’, ‘甘草’, ‘桔梗’, ‘人参’, ‘川芎’, ‘茯苓’, ‘枳壳’, ‘前胡’, ‘羌活’, ‘独活’, ‘生姜’, ‘薄荷’]
仙方活命饮 (Immortal Formula Life-Saving Drink)	BF7	[‘白芷’, ‘贝母’, ‘防风’, ‘赤芍’, ‘当归’, ‘甘草’, ‘皂角刺’, ‘穿山甲’, ‘天花粉’, ‘乳香’, ‘没药’, ‘金银花’, ‘陈皮’]

Basic formula name	Code name	Composition
六安煎 (Liu'an Decoction)	BF8	['陈皮', '半夏', '茯苓', '甘草', '苦杏仁', '芥子', '生姜']
升降散 (Ascending and Descending Powder)	BF9	['僵蚕', '蝉蜕', '姜黄', '大黄']
半夏厚朴汤 (Pinellia and Magnolia Bark Decoction)	BF10	['半夏', '厚朴', '茯苓', '生姜', '紫苏叶']
半夏泻心汤 (Pinellia Decoction for Draining the Heart)	BF11	['甘草', '人参', '黄芩', '干姜', '半夏', '黄连', '大枣']
厚朴生姜半夏甘草人参汤 (Decoction of Magnolia Bark, Ginger, Pinellia, Licorice, and Ginseng)	BF12	['厚朴', '生姜', '半夏', '甘草', '人参']
参附汤 (Ginseng and Aconite Decoction)	BF13	['人参', '附子']
四苓散 (Four-Ingredient Powder with Poria)	BF14	['白术', '猪苓', '茯苓', '桂枝', '泽泻']
四逆加人参汤 (Four Reversals plus Ginseng Decoction)	BF15	['甘草', '干姜', '附子', '人参']
四逆汤 (Four Reversals Decoction)	BF16	['甘草', '干姜', '附子']
地黄丸（六味地黄丸） (Six-Ingredient Rehmannia Pill)	BF17	['熟地黄', '山萸肉', '山药', '泽泻', '牡丹皮', '茯苓']
培元固本散 (Primordial Qi-Tonifying and Root-Strengthening Powder)	BF18	['紫河车', '鹿茸', '人参', '五灵脂', '三七', '琥珀']
外感发热病指南 (Guidelines for Treating Fever due to External Pathogen)	BF19	['羌活', '石膏', '葛根', '柴胡', '白芷', '黄芩', '白芍', '甘草', '桔梗', '生姜', '大枣', '人参']
大黄附子汤 (Rhubarb-Aconite Decoction)	BF20	['大黄', '附子', '细辛']
封髓丹 (Marrow-Sealing Pill)	BF21	['黄柏', '砂仁', '甘草']
射干麻黄汤 (Belamcanda and Ephedra Decoction)	BF22	['射干', '麻黄', '生姜', '细辛', '紫菀', '款冬花', '五味子', '大枣', '半夏']
小半夏加茯苓汤 (Minor Pinellia plus Poria Decoction)	BF23	['半夏', '生姜', '茯苓']
小柴胡汤 (Minor Bupleurum Decoction)	BF24	['柴胡', '黄芩', '人参', '甘草', '生姜', '大枣', '半夏']

Basic formula name	Code name	Composition
小陷胸汤 (Minor Chest Collapse Decoction)	BF25	[‘黄连’, ‘半夏’, ‘瓜蒌’]
小青龙汤 (Minor Blue-Green Dragon Decoction)	BF26	[‘麻黄’, ‘白芍’, ‘细辛’, ‘干姜’, ‘甘草’, ‘桂枝’, ‘五味子’, ‘半夏’]
干姜人参半夏丸 (Dry Ginger, Ginseng, and Pinellia Pill)	BF27	[‘干姜’, ‘人参’, ‘半夏’, ‘生姜’]
干姜黄芩黄连人参汤 (Dry Ginger, Scutellaria, Coptis, and Ginseng Decoction)	BF28	[‘干姜’, ‘黄芩’, ‘黄连’, ‘人参’]
引火汤 (Fire-Attracting Decoction)	BF29	[‘熟地黄’, ‘巴戟天’, ‘天冬’, ‘麦冬’, ‘茯苓’, ‘五味子’, ‘肉桂’]
当归补血汤 (Angelica Blood-Tonifying Decoction)	BF30	[‘黄芪’, ‘当归’]
慢性支气管炎指南 1 (Chronic Bronchitis Guidelines 1)	BF31	[‘麻黄’, ‘蝉蜕’, ‘苦杏仁’, ‘茯苓’, ‘干姜’, ‘细辛’, ‘五味子’, ‘半夏’, ‘生姜’, ‘甘草’, ‘紫菀’, ‘款冬花’, ‘附子’]
慢性支气管炎指南 2 (Chronic Bronchitis Guidelines 2)	BF32	[‘紫河车’, ‘鹿茸’, ‘人参’, ‘蛤蚧’, ‘三七’, ‘琥珀’, ‘贝母’, ‘金蝉花’, ‘沉香’, ‘麻黄’]
排脓汤 (Pus-Expelling Decoction)	BF33	[‘甘草’, ‘桔梗’, ‘生姜’, ‘大枣’]
旋覆代赭汤 (Xuanfu Daizhe Decoction)	BF34	[‘旋覆花’, ‘人参’, ‘生姜’, ‘代赭石’, ‘甘草’, ‘半夏’, ‘大枣’]
术附汤（近效术附汤） (Atractylodes-Aconite Decoction (Modified))	BF35	[‘白术’, ‘附子’, ‘甘草’, ‘生姜’, ‘大枣’]
柴胡桂枝汤 (Bupleurum and Cinnamon Twig Decoction)	BF36	[‘桂枝’, ‘黄芩’, ‘人参’, ‘甘草’, ‘半夏’, ‘白芍’, ‘大枣’, ‘生姜’, ‘柴胡’]
柴葛解肌汤 (Bupleurum and Kudzu Decoction)	BF37	[‘柴胡’, ‘葛根’, ‘甘草’, ‘黄芩’, ‘白芍’, ‘羌活’, ‘白芷’, ‘桔梗’, ‘生姜’, ‘大枣’, ‘石膏’]
栀子甘草豉汤 (Decoction of Gardenia, Licorice, and Fermented Soybean)	BF38	[‘栀子’, ‘甘草’, ‘淡豆豉’]
桂枝加芍药生姜各一两人参三两新加汤 (Modified Cinnamon Twig Decoction with Peony, Ginger, and Ginseng)	BF39	[‘桂枝’, ‘白芍’, ‘甘草’, ‘人参’, ‘生姜’, ‘大枣’]

Basic formula name	Code name	Composition
桂枝加附子汤 (Cinnamon Twig with Aconite Decoction)	BF40	[‘桂枝’, ‘白芍’, ‘甘草’, ‘生姜’, ‘大枣’, ‘附子’]
桂枝加龙骨牡蛎汤 (Cinnamon Twig with Dragon Bone and Oyster Shell Decoction)	BF41	[‘桂枝’, ‘白芍’, ‘生姜’, ‘甘草’, ‘大枣’, ‘龙骨’, ‘牡蛎’]
桂枝去芍药加麻黄细辛附子汤 (Cinnamon Twig Decoction (without Peony) with Ephedra, Asarum, and Aconite)	BF42	[‘桂枝’, ‘生姜’, ‘甘草’, ‘大枣’, ‘麻黄’, ‘细辛’, ‘附子’]
桂枝汤 (Cinnamon Twig Decoction)	BF43	[‘桂枝’, ‘白芍’, ‘甘草’, ‘生姜’, ‘大枣’]
桂枝甘草汤 (Cinnamon Twig and Licorice Decoction)	BF44	[‘桂枝’, ‘甘草’]
桂枝甘草龙骨牡蛎汤 (Cinnamon Twig, Licorice, Dragon Bone, and Oyster Shell Decoction)	BF45	[‘桂枝’, ‘甘草’, ‘牡蛎’, ‘龙骨’]
桂枝附子汤 (Cinnamon Twig and Aconite Decoction)	BF46	[‘桂枝’, ‘附子’, ‘甘草’, ‘生姜’, ‘大枣’]
桔梗汤 (Platycodon Decoction)	BF47	[‘桔梗’, ‘甘草’]
橘皮汤 (Tangerine Peel Decoction)	BF48	[‘陈皮’, ‘生姜’]
止呕验方 (Empirical Formula for Stopping Vomiting)	BF49	[‘半夏’, ‘茯苓’, ‘代赭石’, ‘生姜’, ‘生姜’]
止嗽散 (Antitussive Powder)	BF50	[‘桔梗’, ‘荆芥’, ‘紫菀’, ‘百部’, ‘白前’, ‘甘草’, ‘陈皮’]
泽泻汤 (Alisma Decoction)	BF51	[‘泽泻’, ‘白术’]
活络效灵丹 (Efficacious Pill for Activating the Collaterals)	BF52	[‘当归’, ‘丹参’, ‘乳香’, ‘没药’]
温胆汤 (Warm the Gallbladder Decoction)	BF53	[‘半夏’, ‘竹茹’, ‘枳实’, ‘陈皮’, ‘甘草’, ‘茯苓’]
甘草干姜汤 (Licorice and Dried Ginger Decoction)	BF54	[‘甘草’, ‘干姜’]
甘草干姜茯苓白术汤 (Licorice, Dried Ginger, Poria, and Atractylodes Decoction)	BF55	[‘甘草’, ‘白术’, ‘干姜’, ‘茯苓’]
甘草泻心汤 (Licorice Decoction for Draining the Heart)	BF56	[‘甘草’, ‘黄芩’, ‘干姜’, ‘半夏’, ‘黄连’, ‘大枣’]

Basic formula name	Code name	Composition
甘草附子汤 (Licorice and Aconite Decoction)	BF57	['甘草', '附子', '白术', '桂枝']
生姜泻心汤 (Ginger Decoction for Draining the Heart)	BF58	['生姜', '甘草', '人参', '黄芩', '干姜', '半夏', '黄连', '大枣']
白术汤（四君子汤）(Four Gentlemen Decoction)	BF59	['白术', '茯苓', '人参', '甘草']
肺胀病指南-缓解期 (Guidelines for Lung Distension Disease - Remission Phase)	BF60	['鹿茸', '三七', '琥珀', '紫河车', '蛤蚧', '人参', '沉香', '贝母', '灵芝', '鱼肚', '冬虫夏草']
胃痞病指南 1 (Guidelines for Gastric Distention Disorder 1)	BF61	['升麻', '紫菀', '紫苏梗', '枇杷叶', '半夏', '厚朴', '党参', '甘草']
芍药甘草汤 (Peony and Licorice Decoction)	BF62	['白芍', '甘草']
芍药甘草附子汤 (Peony, Licorice, and Aconite Decoction)	BF63	['白芍', '甘草', '附子']
芎🄵汤（佛手散）(Fo Shou Powder (Chuanxiong Decoction))	BF64	['当归', '川芎']
苓甘五味加姜辛半夏杏仁汤 (Ling Gan Wu Wei Decoction with Added Ginger, Asarum, Pinellia, and Apricot Kernel)	BF65	['茯苓', '甘草', '五味子', '干姜', '细辛', '半夏', '苦杏仁']
苓甘五味姜辛汤 (Ling Gan Wu Wei Decoction with Ginger and Asarum)	BF66	['茯苓', '甘草', '干姜', '细辛', '五味子']
茯苓四逆汤 (Poria Four Reversals Decoction)	BF67	['茯苓', '人参', '附子', '甘草', '干姜']
茯苓杏仁甘草汤 (Poria, Apricot Kernel, and Licorice Decoction)	BF68	['茯苓', '苦杏仁', '甘草']
茯苓甘草汤 (Poria and Licorice Decoction)	BF69	['茯苓', '桂枝', '甘草', '生姜']
贞元饮 (Zhen Yuan Drink)	BF70	['熟地黄', '甘草', '当归']
越婢汤 (Yue Bi Decoction)	BF71	['麻黄', '石膏', '生姜', '大枣', '甘草']
附子理中汤 (Aconite Decoction to Regulate the Middle)	BF72	['附子', '人参', '干姜', '甘草', '白术']
麻黄加术汤 (Ephedra plus Atractylodes Decoction)	BF73	['麻黄', '桂枝', '甘草', '苦杏仁', '白术']

Basic formula name	**Code name**	**Composition**
麻黄汤 (Ephedra Decoction)	BF74	[‘麻黄’, ‘桂枝’, ‘甘草’, ‘苦杏仁’]
麻黄细辛附子汤 (Ephedra, Asarum, and Aconite Decoction)	BF75	[‘麻黄’, ‘细辛’, ‘附子’]
麻黄附子甘草汤 (Ephedra, Aconite, and Licorice Decoction)	BF76	[‘麻黄’, ‘甘草’, ‘附子’]
黄芩汤 (Scutellaria Decoction)	BF77	[‘黄芩’, ‘白芍’, ‘甘草’, ‘大枣’]
黄芪汤 (Astragalus Decoction)	BF78	[‘黄芪’, ‘人参’, ‘甘草’]
鼻渊病-虚证 (Nasal Catarrh - Deficiency Syndrome)	BF79	[‘柴胡’, ‘野菊花’, ‘蔓荆子’, ‘黄芩’, ‘辛夷’, ‘白芷’, ‘鱼腥草’, ‘薏苡仁’, ‘地龙’, ‘蒲公英’, ‘桔梗’, ‘甘草’]

Table 1: Basic formula mentioned at the main passage.

2.2 Incremental pre-training

Qwen2.5-7B-Instruct				
Parameter	Indicators			
warmup_ratio	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
0.3	1.5539	7.0727	2.6177	5.8377
0.25	1.4814	7.3579	2.7163	6.0129
0.2	1.9334	7.6204	3.3736	5.9013
0.15	2.5272	10.6926	4.3944	7.9225
0.1	1.9807	8.5531	3.0597	7.3205
0.05	2.1963	9.4712	3.0695	8.3924
lr				
5E-05	2.5272	10.6926	4.3944	7.9225
5E-06	3.418	11.8297	4.9176	9.8096
1E-05	4.7848	14.7422	6.3144	12.5815
1E-06	2.8119	10.3312	4.255	8.4434

Table 2: Detailed results after ipt on Qwen2.5-7B-Instruct.

gemma-2-9b-it				
Parameter	Indicators			
warmup_ratio	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
0.3	2.3695	8.7506	3.9212	6.8693
0.25	1.2191	5.44	2.2982	3.83
0.2	0.4949	2.2609	0.8643	1.6509
0.15	1.3657	5.9373	2.1652	4.4843
0.1	1.5328	7.1989	2.6361	5.7224
0.05	0.7569	3.7203	1.1636	2.6861
lr				
5E-05	2.3695	8.7506	3.9212	6.8693
5E-06	1.3439	7.7014	2.9662	5.0507
1E-05	2.2995	9.8494	3.9622	7.3681
1E-06	0.8566	5.463	2.1043	3.4758

Table 3: Detailed results after ipt on gemma-2-9b-it.

vicuna-7b-v1.5				
Parameter	Indicators			
warmup_ratio	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
0.3	1.8091	6.4451	3.244	4.8857
0.25	2.3937	8.7567	4.1955	6.5575
0.2	1.9571	8.4492	4.0412	5.7963
0.15	1.8733	6.2361	3.2174	4.9087
0.1	2.0425	6.7432	3.2768	5.5039
0.05	1.7417	6.6945	3.043	5.0585
lr				
5E-05	2.3937	8.7567	4.1955	6.5575
5E-06	0.79	3.3161	1.7364	2.3397
1E-05	0.9948	4.1127	1.7211	3.4587
1E-06	0.456	3.665	1.8572	1.8912

Table 4: Detailed results after ipt on vicuna-7b-v1.5.

Meta-Llama-3-8B-Instruct				
Parameter	Indicators			
warmup_ratio	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
0.3	0.5562001	4.2050252	1.7990472	2.0809437
0.25	0.5368045	4.2427926	1.7827917	2.0351105
0.2	0.5402812	4.2470934	1.7738455	2.0484962
0.15	0.5636217	4.268132	1.871562	2.0407951
0.1	0.5187724	4.159164	1.7095902	2.0443288
0.05	0.5118308	4.185552	1.7008852	2.0480913
lr				
5E-05	0.5461897	4.2638433	1.7970561	2.0536855
1E-05	1.903763	8.6379927	3.6411147	6.8700554
5E-06	1.5222739	8.2123328	3.2328271	5.8724195
1E-06	0.1901947	2.8270396	0.7306931	1.3044748
5E-07	0.0444459	1.778219	0.2611099	0.6808455

Table 5: Detailed results after ipt on Meta-Llama-3-8B-Instruct.

Mistral-7B-Instruct-v0.3				
Parameter	Indicators			
warmup_ratio	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
0.3	0.6745881	9.5981312	4.3522113	2.6141735
0.25	0.9911208	12.0798838	4.9046659	4.1456815
0.2	1.9887954	10.6097794	4.2397013	7.4887342
0.15	1.5105443	7.8448934	3.2824601	6.3556967
0.1	1.4472153	8.0829003	3.3280991	6.2898209
0.05	0.3960097	4.7188696	1.4317643	2.3020964
lr				
5E-05	0.6745881	9.5981312	4.3522113	2.6141735
1E-05	0.9911208	12.0798838	4.9046659	4.1456815
5E-06	1.9887954	10.6097794	4.2397013	7.4887342
1E-06	1.5105443	7.8448934	3.2824601	6.3556967
5E-07	1.4472153	8.0829003	3.3280991	6.2898209

Table 6: Detailed results after ipt on Mistral-7B-Instruct-v0.3.

DeepSeek-v2-Lite				
Parameters	Indicators			
warmup ratio	BLEU-4	rough-1	rough-2	rough-l
0.45	0.4099	5.2387	2.3509	1.6119
0.4	0.3677	4.6677	1.9344	1.5238
0.35	0.4225	4.3764	1.8895	1.6375
0.3	0.393	5.7237	2.3706	1.6889
0.25	0.3885	4.3819	1.8703	1.598
0.2	0.4071	5.7427	2.4126	1.6681
0.1	0.3629	4.6398	1.826	1.5765
lr				
5.00E-04	0.3067	12.5267	5.7375	1.3504
5.00E-05	0.4071	5.7427	2.4126	1.6681
1.00E-05	0.3579	5.4836	2.1809	1.5944
5.00E-06	0.2852	6.1879	1.5944	1.4578

Table 7: Detailed results after ipt on DeepSeek-v2-Lite.

2.3 Supervised fine-tuning

Qwen2.5-7B-Instruct_HGipt_HGsftD0					
Test Dataset	Parameters	Indicators			
HGtestD0/1/2	warmup_ratio	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-l
	0.3	38.7187675	66.0590083	41.692556	62.9205221
	0.25	40.4950521	67.4783502	43.355744	64.6198823
	0.2	39.6102508	66.7687402	42.9161276	63.9563686
	0.15	41.2578223	67.9112988	44.4640174	65.2876414
	0.1	39.491412	66.6576238	42.27014	63.7761647
	0.05	38.8057366	66.5424703	41.9761484	63.8676223
	lr				
	5.00E-05	39.2827237	66.3411672	41.8282225	63.700541
	1.00E-05	39.3274586	66.7601047	42.3691453	63.880712
	5.00E-06	36.4683189	63.6194952	39.1059481	60.5825124
	1.00E-06	19.8026645	36.579551	20.6831843	34.4974148
HGtestD3/9	Best parameters above	37.0107359	63.0977105	39.4052953	60.8653502

Table 8: Detailed results of Qwen2.5-7B-Instruct_HGipt_HGsftD0.

Qwen2.5-7B-Instruct_HGipt_HGsftD1					
Test Dataset	Parameters	Indicators			
HGtestD0/1/2	warmup_ratio	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
	0.3	42.3901859	69.6307057	45.5700159	66.808897
	0.25	42.5150978	69.2234283	45.7488453	66.4650535
	0.2	33.7580436	60.1575738	34.7021135	57.9491062
	0.15	41.973619	68.733106	45.1957384	66.1256687
	0.1	36.3513029	63.1103135	37.4022453	61.884063
	0.05	43.2050852	69.7137142	46.4252709	67.0456886
	lr				
	5.00E-05	42.4170604	69.1523256	46.0816622	66.3159541
	1.00E-05	40.3597042	67.5974385	43.5525184	64.4390378
	5.00E-06	35.1564181	63.4894513	38.101422	59.6938012
	1.00E-06	18.7231657	34.9939801	19.2638971	33.0946686
HGtestD3/9	Best parameters above	37.7504978	63.47429	40.0614461	61.2746877

Table 9: Detailed results of Qwen2.5-7B-Instruct_HGipt_HGsftD1.

Qwen2.5-7B-Instruct_HGipt_HGsftD2					
Test Dataset	Parameters	Indicators			
HGtestD0/1/2	warmup_ratio	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
	0.3	42.2076435	68.8410054	45.8005741	66.4229557
	0.25	42.6819878	69.1485861	46.3581957	66.6484501
	0.2	41.0717797	68.7103943	44.8610145	65.9857116
	0.15	24.8965185	49.4228923	24.0955694	48.2410782
	0.1	40.9372767	68.0116002	44.2538503	65.1744787
	0.05	38.9564385	66.3832342	42.1023737	63.2853908
	lr				
	5.00E-05	41.9781898	68.5158081	44.6731591	66.2149768
	1.00E-05	37.3090703	65.0287676	39.8561809	62.3193764
	5.00E-06	35.6483761	62.7688173	38.4166617	59.7114424
	1.00E-06	17.8145666	34.2123841	18.8196568	32.2963648
HGtestD3/9	Best parameters above	38.0107359	64.0575086	40.6640129	61.1658073

Table 10: Detailed results of Qwen2.5-7B-Instruct_HGipt_HGsftD2.

Qwen2.5-72B-Instruct_HGsftD0				
Parameters	Indicators			
warmup_ratio	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
0.3	37.3939	63.0763	39.5584	60.1393
0.25	37.8639	62.6368	39.9763	60.0550
0.2	38.3424	64.3453	40.7453	61.6372
0.15	37.2361	64.0411	39.9468	61.0353
0.1	38.6435	65.2027	41.1638	62.2795
0.05	38.6953534	65.3209	41.7783	62.4437
lr				
5E-05	38.6953534	65.3209	41.7783	62.4437
5E-06	14.2762958	28.8872	10.1943	26.4156876
1E-05	27.5753	54.4213	29.2553	51.3260
1E-06	0.3791	5.5431	0.6497	2.5154

Table 11: Detailed results of Qwen2.5-72B-Instruct_HGsftD0.

Qwen2.5-72B-Instruct_HGsftD1				
Parameters	Indicators			
warmup_ratio	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
0.3	37.5456	64.7617	40.0752	62.0447
0.25	36.3860	63.0301	38.9314	60.4980
0.2	37.5404	64.2354	40.2079	61.7415
0.15	37.7255	64.4571	39.8339	61.8731
0.1	37.7496	64.3854	40.3635	61.5542
0.05	37.2886	64.2787	40.1926	61.5033
lr				
5E-05	37.5456	64.7617	40.0752	62.0447
5E-06	10.5435	23.1965	8.0433	20.5759
1E-05	23.9170	47.1621	25.5040	43.4555
1E-06	0.31258	5.3537	0.4657	2.4048

Table 12: Detailed results of Qwen2.5-72B-Instruct_HGsftD1.

Qwen2.5-72B-Instruct_HGsftD2				
Parameters	Indicators			
warmup_ratio	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
0.3	36.0842514	63.5726607	38.0404147	60.6307589
0.25	35.9250	63.0736	38.2122	60.4369
0.2	37.0951	63.8120	39.1827	61.1893
0.15	35.3839	62.4127	37.1774	59.8589
0.1	35.2154	62.4457	37.3641	59.8311
0.05	36.0065	62.8238	38.2779	60.1829
lr				
5E-05	37.0951	63.8120	39.1827	61.1893
5E-06	13.7676	26.6240	13.0402	24.3752
1E-05	27.036999	52.3989	27.3273	49.6552
1E-06	0.3496	5.8441	0.5696	2.7262

Table 13: Detailed results of Qwen2.5-72B-Instruct_HGsftD2.

2.4 Data ablation study results

Qwen2.5-7B-Instruct_HGipt_HGsftD0_woBI				
Parameters	Indicators			
warmup_ratio	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
0.3	41.8308172	65.2451327	42.2254034	61.1282289
0.25	41.492495	64.665915	42.0982499	60.4577922
0.2	42.1370628	64.9060174	42.4684687	60.8638999
0.15	40.5971624	64.2806458	40.4271442	60.2519397
0.1	42.3221386	65.6029228	43.18773	61.3731112
0.05	42.4646457	66.4327084	43.4742648	62.6026791
lr				
5.00E-05	42.4646457	66.4327084	43.4742648	62.6026791
1.00E-05	39.4302196	62.7735882	38.9195229	59.0611828
5.00E-06	37.878748	61.7332446	37.5276137	58.366592
1.00E-06	34.2343784	58.1339445	33.4264878	54.4525343

Table 14: Detailed results of Qwen2.5-7B-Instruct_HGipt_HGsftD0_woBI.

Qwen2.5-7B-Instruct_HGipt_HGsftD1_woBI				
Parameters	Indicators			
warmup_ratio	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
0.15	38.076062	63.010048	39.0708366	59.9461669
0.2	36.9567659	61.7070733	37.2501925	58.4122175
0.25	35.8545737	59.3269358	35.7013294	56.5840632
0.3	37.9788296	62.0691798	38.4833673	59.2383361
0.05	35.6195958	59.9840938	35.2122249	56.7755181
0.1	36.4359718	60.370542	36.5157866	57.3520441
lr				
5.00E-06	32.9654491	57.7016267	32.893225	53.9952861
1.00E-06	20.8666069	37.940449	19.5920579	35.3314639
5.00E-05	38.076062	63.010048	39.0708366	59.9461669
1.00E-05	35.6195958	59.9840938	35.2122249	56.7755181

Table 15: Detailed results of Qwen2.5-7B-Instruct_HGipt_HGsftD1_woBI.

Qwen2.5-7B-Instruct_HGipt_HGsftD2_woBI				
Parameters	Indicators			
warmup_ratio	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
0.2	38.0014867	62.5717996	39.0536152	59.5910315
0.25	37.7373044	62.4508969	38.5688319	59.2676717
0.3	38.0864969	62.7937308	39.4479166	59.7578906
0.15	39.651038	63.8367773	40.9607723	60.6123294
0.05	39.0132652	63.1456225	40.2161531	60.2761807
0.1	39.1858664	63.484385	39.6915625	60.2318417
lr				
1.00E-06	26.6016167	46.5481899	25.2145134	43.5350873
1.00E-05	35.8100661	60.635529	35.381531	57.353302
5.00E-05	39.651038	63.8367773	40.9607723	60.6123294
5.00E-06	34.7290319	58.8836276	33.6888112	55.7455241

Table 16: Detailed results of Qwen2.5-7B-Instruct_HGipt_HGsftD2_woBI.

Qwen2.5-7B-Instruct_HGipt_HGsftD0_woBF				
Parameters	Indicators			
warmup_ratio	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
0.25	35.1249383	50.7267239	35.5299116	50.2220144
0.1	35.5622171	51.005279	35.7582695	50.3402032
0.15	34.7457605	50.3341972	34.8837741	49.843259
0.3	34.9231275	49.5614773	34.9989287	49.4111424
0.05	31.2951802	48.2954509	31.658174	47.9395073
0.2	34.1228899	49.8346555	34.276867	49.572316
lr				
1.00E-06	28.9016178	46.5111379	28.6905616	45.8087859
5.00E-05	35.5622171	51.005279	35.7582695	50.3402032
1.00E-05	33.0878808	50.2531323	33.6932108	49.6309884
5.00E-06	31.8757167	48.1735577	32.0065364	47.8640682

Table 17: Detailed results of Qwen2.5-7B-Instruct_HGipt_HGsftD0_woBF.

Qwen2.5-7B-Instruct_HGipt_HGsftD1_woBF				
Parameters	Indicators			
warmup_ratio	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
0.3	28.8565246	45.8519928	29.1942948	45.3896125
0.25	29.2703711	45.3349164	29.5493196	44.9669414
0.15	29.9026997	47.0107083	30.6274059	46.579964
0.1	29.329638	45.248956	29.0281236	45.1931829
0.2	29.3738304	47.2218488	30.0325905	46.7216954
0.05	30.1566413	47.4615746	30.0184171	47.071858
lr				
5.00E-05	29.9026997	47.0107083	30.6274059	46.579964
1.00E-06	23.4142451	38.9300493	23.0617588	38.1086257
5.00E-06	29.5746441	46.8186393	29.231917	46.2625296
1.00E-05	30.1566413	47.4615746	30.0184171	47.071858

Table 18: Detailed results of Qwen2.5-7B-Instruct_HGipt_HGsftD1_woBF.

Qwen2.5-7B-Instruct_HGipt_HGsftD2_woBF				
Parameters	Indicators			
warmup_ratio	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
0.3	29.4144082	47.8231394	29.7651536	47.2416278
0.25	30.7549606	47.7115361	31.2863536	47.3319723
0.2	29.8138032	47.5044428	30.2736362	47.146139
0.1	29.2337103	46.3027691	29.1788031	45.7876974
0.15	30.2253142	47.8412365	30.2447247	46.9535512
0.05	30.6307549	47.4696407	30.6431349	47.0372348
lr				
1.00E-06	22.2073141	38.2372957	22.2696382	37.3306303
5.00E-05	30.7549606	47.7115361	31.2863536	47.3319723
5.00E-06	29.0602114	46.4998727	28.9367688	46.0277749
1.00E-05	30.6307549	47.4696407	30.6431349	47.0372348

Table 19: Detailed results of Qwen2.5-7B-Instruct_HGipt_HGsftD2_woBF.

2.5 Task ablation study results

Qwen2.5-7B-Instruct_HGipt_HGsftD0_woBD				
Parameters	Indicators			
warmup_ratio	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
0.3	38.0775974	56.8384373	39.7108605	55.5079858
0.25	37.2770619	57.1727582	39.0378967	55.3550568
0.2	36.4958206	56.0041194	38.0423915	54.3156197
0.15	38.2842496	57.5964084	40.1279574	56.0999331
0.1	37.8250357	57.083826	39.9458819	55.5714518
0.05	37.5746085	57.0843373	39.5828902	55.6704748
lr				
5.00E-05	38.5770959	57.5509707	40.5783753	56.1993903
1.00E-05	35.6237044	55.7064618	36.51224	53.6510413
5.00E-06	35.3434154	55.5739616	37.3515688	53.5124668
1.00E-06	22.4375769	39.3657943	22.7106528	37.8721846

Table 20: Detailed results of Qwen2.5-7B-Instruct_HGipt_HGsftD0_woBD.

Qwen2.5-7B-Instruct_HGipt_HGsftD0_woDP				
Parameters	Indicators			
warmup_ratio	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
0.3	38.9577343	65.8298512	42.1890165	62.9269265
0.25	38.6176873	66.0879163	42.1324919	62.9237624
0.2	38.185345	64.9763948	41.243617	62.1159878
0.15	35.4501303	62.0671628	39.0439312	58.908696
0.1	21.8020665	49.3177525	24.8140258	43.0899462
0.05	38.9768422	65.4442181	42.0804594	62.2611249
lr				
5.00E-05	38.6308777	65.3124689	41.5452448	62.3880434
1.00E-05	35.7422418	63.395528	38.7505823	60.1686453
5.00E-06	32.3411686	59.5918079	34.9083152	56.3586343
1.00E-06	18.5745437	35.9409721	19.3270624	33.941192

Table 21: Detailed results of Qwen2.5-7B-Instruct_HGipt_HGsftD0_woDP.

Qwen2.5-7B-Instruct_HGipt_HGsftD0_woCP				
Parameters	Indicators			
warmup_ratio	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
0.3	33.0278076	58.2747999	36.7559884	54.4973521
0.25	31.7611806	57.452358	35.4371284	53.6257476
0.2	32.1316716	57.7458686	35.1660344	54.261749
0.15	32.6017296	57.8212801	36.3556831	54.1275459
0.1	31.9322071	57.3803911	35.3579229	53.8672516
0.05	31.9981598	57.7563179	35.0693959	53.7867561
lr				
5.0e-5	33.6190593	58.9508174	36.8237379	54.9986136
1.00E-05	29.7966862	55.333856	33.3481985	51.5143226
5.00E-06	28.8475585	54.3843659	32.5279096	50.3207947
1.00E-06	21.511766	39.0756838	22.628069	37.0718585

Table 22: Detailed results of Qwen2.5-7B-Instruct_HGipt_HGsftD0_woCP.

Qwen2.5-7B-Instruct_HGipt_HGsftD0_woBR				
Parameters	Indicators			
warmup_ratio	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
0.3	38.2740558	61.7638842	37.538149	58.5007137
0.25	34.7067869	58.8085811	34.4249797	55.1054518
0.2	38.0438162	62.0998328	37.4273239	58.6246049
0.15	38.0007496	62.2205736	37.4946469	58.8062554
0.1	38.5203572	62.6648928	38.3779897	59.1973851
0.05	38.1484311	62.2454096	38.0252727	58.8616098
lr				
5.00E-05	37.5217762	60.9010385	36.8112775	57.5505119
1.00E-05	35.4994629	60.0717022	34.9390984	56.7188812
5.00E-06	33.2137713	57.9016843	32.4752528	54.4966519
1.00E-06	17.9584432	33.345722	17.1381257	31.2334002

Table 23: Detailed results of Qwen2.5-7B-Instruct_HGipt_HGsftD0_woBR.

2.6 Analysis on generalization ability across different schools

Model	Parameters	Indicators			
	warmup_ratio	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
Qwen2.5-7B-Instruct	\	0.1752	3.3736	0.0273	1.382
Qwen2.5-7B-Instruct_HGipt	Best parameters	5.5577	20.554	1.3917	16.7598
Qwen2.5-7B-Instruct_HGipt_HGsftD0	Best parameters	13.1128	30.94	7.0155	27.9685
Qwen2.5-7B-Instruct_HGipt_HGsftD1	Best parameters	11.9633	28.125	6.0625	24.9039
Qwen2.5-7B-Instruct_HGipt_HGsftD2	Best parameters	13.0696	29.4794	7.0907	26.2874
Qwen2.5-7B-Instruct_HGipt_GAMsft	0.15	21.5775699	40.3737655	17.5733367	39.163015
	0.3	21.6235135	40.083271	17.4907055	39.0410741
	0.1	21.9688061	40.5297135	18.4924038	39.5927199
	0.25	20.9586091	39.0761609	17.038278	38.1740623
	0.05	21.8136269	40.6042634	17.9623007	39.3493823
	0.2	22.8009672	41.2240172	19.344534	40.0746791
	lr				
	5.00E-05	17.7470703	38.1171295	8.8238887	34.7576062
	5.00E-06	22.8009672	41.2240172	19.344534	40.0746791
	1.00E-05	21.4036419	40.5632622	15.8058282	38.7455593
	1.00E-06	18.6208187	35.3275216	17.6342402	33.9340184

Table 24: Detailed results on GAMtest.

2.7 Analysis on the impact of CoT

Qwen2.5-7B-Instruct_HGipt_HGsftD3				
Parameters	Indicators			
warmup_ratio	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
0.2	33.2864467	56.9874342	33.2533871	53.2617399
0.05	32.3331154	54.5593156	32.3253985	51.0600889
0.25	34.094	57.2148179	34.0538532	53.401774
0.1	33.9140993	57.2566612	33.9928099	53.6298516
0.15	34.1108333	56.4819374	34.0058454	52.6668331
0.3	33.3497937	56.1346787	32.9678537	52.589296
lr				
1.00E-06	27.4166329	50.220768	26.9230687	46.4890986
5.00E-05	34.1108333	56.4819374	34.0058454	52.6668331
1.00E-05	27.6223615	49.8916871	25.423988	47.2372976
5.00E-06	30.5999421	53.9960594	29.86377	50.3183295

Table 25: Detailed results of Qwen2.5-7B-Instruct_HGipt_HGsftD3 on HGtestD0/1/2.

Qwen2.5-7B-Instruct_HGipt_HGsftD3_cot				
Parameters	Indicators			
warmup_ratio	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
0.05	38.9391951	65.32016	42.1523476	62.1491958
0.2	38.3054422	64.8921224	41.1339286	61.619022
0.15	38.1116386	64.668274	41.2156097	61.5937211
0.25	39.3165875	65.5147049	42.2507161	62.5167441
0.1	38.8431627	65.3640016	41.9830412	62.2531931
0.3	38.1165761	64.8240477	41.324689	61.832989
lr				
5.00E-05	39.3165875	65.5147049	42.2507161	62.5167441
5.00E-06	37.1419492	63.7941686	39.2375601	61.1838755
1.00E-06	34.2878705	60.3839115	36.206847	57.5181245
1.00E-05	38.7365802	65.0413478	41.2832912	62.0889169

Table 26: Detailed results of Qwen2.5-7B-Instruct_HGipt_HGsftD3_cot on HGtestD0/1/2.

Qwen2.5-7B-Instruct_HGipt_HGsftD3_cot				
Parameters	Indicators			
warmup_ratio	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
0.05	42.5294121	63.8058158	43.0409138	60.150076
0.3	42.9332082	64.7902055	43.0121326	60.9946315
0.15	43.3265763	65.361966	43.5024538	61.0257469
0.2	43.7318197	66.0722035	44.3558075	62.3535846
0.1	42.7953764	64.9571747	43.4925228	61.3507126
0.25	43.9606256	66.3111111	44.9327746	62.1755405
lr				
5.00E-05	43.9606256	66.3111111	44.9327746	62.1755405
1.00E-05	35.1925065	57.5473532	33.6711278	54.4996784
5.00E-06	36.5545032	60.3383974	36.5331223	56.4829714
1.00E-06	32.2545346	57.1505947	32.1325432	53.3466664

Table 27: Detailed results of Qwen2.5-7B-Instruct_HGipt_HGsftD3 on HGtestD3/9.

Qwen2.5-7B-Instruct_HGipt_HGsftD3_cot				
Parameters	Indicators			
warmup_ratio	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
0.05	42.5294121	63.8058158	43.0409138	60.150076
0.3	42.9332082	64.7902055	43.0121326	60.9946315
0.15	43.3265763	65.361966	43.5024538	61.0257469
0.2	43.7318197	66.0722035	44.3558075	62.3535846
0.1	42.7953764	64.9571747	43.4925228	61.3507126
0.25	43.9606256	66.3111111	44.9327746	62.1755405
lr				
5.00E-05	43.9606256	66.3111111	44.9327746	62.1755405
1.00E-05	35.1925065	57.5473532	33.6711278	54.4996784
5.00E-06	36.5545032	60.3383974	36.5331223	56.4829714
1.00E-06	32.2545346	57.1505947	32.1325432	53.3466664

Table 28: Detailed results of Qwen2.5-7B-Instruct_HGipt_HGsftD3_cot on HGtestD3/9.

3 Statistical Validation

3.1 Methods

To rigorously assess the reliability and significance of our reported performance metrics, we conducted a comprehensive statistical validation on the fixed test sets. This analysis primarily employed the non-parametric bootstrap method, which is well-suited for estimating the sampling distribution of metrics like BLEU-4 and ROUGE when the underlying distribution is unknown.

Our statistical validation consisted of two main components: (1) constructing confidence intervals (CIs) for the performance of individual models, and (2) performing significance tests to evaluate the impact of specific ablations and enhancements.

It is important to note that we did not perform statistical tests on the results from the incremental pre-training stage. The primary reason is that the performance of all models at this stage was uniformly low (as shown in Section III-D of the main paper), indicating a fundamental lack of task-specific capability before supervised fine-tuning. Since this stage serves as a foundational knowledge injection step rather than a final, deployable model, the focus of our statistical validation was placed on the fine-tuned models and their comparative ablations, which represent the core contributions and claims of this work.

First, for key models evaluated on the test sets HGtestD0/1/2, HGtestD3/9, and GAMtest, we computed 95% percentile bootstrap confidence intervals for their BLEU-4/ROUGE scores. This involved generating $B = 1,000$ bootstrap samples by resampling the test set with replacement and recalculating the BLEU-4 score for each sample. The 95% CI was then derived from the 2.5th and 97.5th percentiles of the resulting empirical distribution of scores.

Second, to quantify the statistical significance of observed performance differences, we performed paired bootstrap tests ($B = 1,000$) for the following critical comparisons:

- Impact of Basic Formulas: Comparing the model fine-tuned on HGsftD0 against its ablated version without basic formulas (HGsftD0_woBF) on the HGtestD0/1/2 test set.
- Impact of Prescription Completion Task: Comparing the model fine-tuned on HGsftD0 against its ablated version without the prescription completion task (HGsftD0_woCP) on the HGtestD0/1/2 test set.
- Impact of Chain-of-Thought (CoT): Comparing the model fine-tuned on HGsftD3 against the model fine-tuned on the CoT-augmented dataset (HGsftD3_cot) on the HGtestD3/9 test set.

For each paired comparison, we calculated the difference in scores (e.g., ΔBLEU-4 = $\text{Score}_{\text{experiment}} - \text{Score}_{\text{control}}$) for each of the 1,000 bootstrap samples. Based on our hypothesis that the ablation would degrade performance (or CoT would improve it), we computed a *one-sided* p-value as $p \approx \Pr(\Delta 0)$ for ablations (where lower score indicates worse performance) and $p \approx \Pr(\Delta 0)$ for CoT enhancement (where higher score indicates better performance). A p-value below 0.05 was considered statistically significant. All statistical analyses were implemented in Python using standard scientific computing libraries.

Finally, to provide qualitative insights into model behavior, we conducted an error analysis on the outputs of the models finetuned on HGsftD0, HGsftD1, and HGsftD2 for Task 3 (prescription generation) on the HGtestD0/1/2 test set. We selected the 3 samples with the lowest BLEU-4 scores as representative failure cases and the 3 samples with the highest BLEU-4 scores as high-quality examples. For each case, we manually examined the input, reference prescription, and model output to identify common error patterns such as missing herbs or redundant herbs.

3.2 Confidence Interval

Table 29: 95% Bootstrap Confidence Intervals for BLEU-4 and ROUGE-1 Across Different Test Sets. All intervals are computed using $B = 1,000$ bootstrap resamples. Values are reported with 4 decimal places.

Model / Condition	**BLEU-4**	**ROUGE-1**	**Test Set**
Fine-tuned Models on HGtestD0/1/2			
HGsftD0	41.4021 – 44.9646	65.0737 – 68.3730	HGtestD0/1/2
HGsftD1	39.9516 – 43.4927	67.1835 – 70.2372	HGtestD0/1/2
HGsftD2	39.6597 – 43.2097	66.7011 – 69.7763	HGtestD0/1/2
Data Ablation			
HGsftD0_woBI	41.3457 – 44.5670	65.2708 – 68.0617	HGtestD0/1/2
HGsftD0_woBF	34.2747 – 36.8497	48.6492 – 50.1451	HGtestD0/1/2
HGsftD1_woBI	37.1162 – 40.2020	61.8848 – 64.6565	HGtestD0/1/2
HGsftD1_woBF	29.1481 – 32.0559	47.4360 – 50.1921	HGtestD0/1/2
HGsftD2_woBI	38.7171 – 41.8076	62.7648 – 65.5761	HGtestD0/1/2
HGsftD2_woBF	29.9803 – 33.0751	48.1642 – 50.9417	HGtestD0/1/2
Task Ablation			

Continued on next page

Table 29 – continued from previous page

Model / Condition	BLEU-4	ROUGE-1	Test Set
HGsftD0_woCP	36.6923 – 39.7131	60.4011 – 62.8150	HGtestD0/1/2
HGsftD0_woBR	41.6731 – 44.8332	63.8374 – 66.6925	HGtestD0/1/2
HGsftD0_woBD	37.8487 – 40.9695	58.6411 – 61.3587	HGtestD0/1/2
HGsftD0_woDP	37.8142 – 41.0804	64.7803 – 67.8058	HGtestD0/1/2
Performance on GAMtest			
Qwen2.5-7B-Instruct	0.1648 – 0.1794	3.1374 – 3.3173	GAMtest
HGipt	4.7935 – 6.4362	19.2259 – 21.9497	GAMtest
HGsftD0	11.6099 – 13.7486	28.3936 – 31.0948	GAMtest
HGsftD1	10.8162 – 13.1455	26.2321 – 28.7781	GAMtest
HGsftD2	11.8182 – 14.1126	27.1356 – 29.7171	GAMtest
GAMsft	21.8523 – 23.7495	40.2512 – 42.1968	GAMtest
Chain-of-Thought (CoT) Enhancement			
HGsftD3 (Non-CoT)	43.9575 – 47.3624	65.7331 – 68.5068	HGtestD3/9
HGsftD3_cot (CoT)	49.1867 – 52.5929	74.0338 – 76.6877	HGtestD3/9
HGsftD3 (Non-CoT)	32.8948 – 36.1570	54.9704 – 57.9285	HGtestD0/1/2
HGsftD3_cot (CoT)	38.6923 – 41.8926	65.1324 – 67.8576	HGtestD0/1/2

Table 30: 95% Bootstrap Confidence Intervals for ROUGE-2 and ROUGE-L Across Different Test Sets. All intervals are computed using $B = 1,000$ bootstrap resamples. Values are reported with 4 decimal places.

Model / Condition	ROUGE-2	ROUGE-L	Test Set
Fine-tuned Models on HGtestD0/1/2			
HGsftD0	41.0348 – 45.8229	63.0159 – 66.1026	HGtestD0/1/2
HGsftD1	43.4312 – 47.7436	64.9265 – 67.9027	HGtestD0/1/2
HGsftD2	43.4810 – 47.9275	64.6969 – 67.6528	HGtestD0/1/2
Data Ablation			
HGsftD0_woBI	42.3633 – 46.4088	61.4946 – 64.2724	HGtestD0/1/2
HGsftD0_woBF	34.4171 – 37.0995	49.4901 – 51.1903	HGtestD0/1/2
HGsftD1_woBI	37.6954 – 41.6269	59.0507 – 61.7323	HGtestD0/1/2

Continued on next page

Table 30 – continued from previous page

Model / Condition	ROUGE-2	ROUGE-L	Test Set
HGsftD1_woBF	29.0539 – 32.9334	47.2266 – 49.9148	HGtestD0/1/2
HGsftD2_woBI	39.5655 – 43.5623	59.6932 – 62.4012	HGtestD0/1/2
HGsftD2_woBF	30.5472 – 34.6152	47.8592 – 50.6115	HGtestD0/1/2
Task Ablation			
HGsftD0_woCP	40.6803 – 43.8858	56.6724 – 58.9358	HGtestD0/1/2
HGsftD0_woBR	42.0191 – 46.0560	60.2990 – 63.0576	HGtestD0/1/2
HGsftD0_woBD	40.1670 – 44.1235	57.2251 – 59.8437	HGtestD0/1/2
HGsftD0_woDP	41.0970 – 45.2446	62.1005 – 65.0212	HGtestD0/1/2
Performance on GAMtest			
Qwen2.5-7B-Instruct	0.0186 – 0.0438	1.3155 – 1.3932	GAMtest
HGipt	0.7469 – 2.1282	15.4881 – 18.1592	GAMtest
HGsftD0	5.2486 – 7.6817	25.6452 – 28.2905	GAMtest
HGsftD1	5.1665 – 7.5055	23.3088 – 25.7370	GAMtest
HGsftD2	6.1414 – 8.6682	24.2915 – 26.7146	GAMtest
GAMsft	18.4017 – 20.2873	39.1021 – 41.0471	GAMtest
Chain-of-Thought (CoT) Enhancement			
HGsftD3 (Non-CoT)	44.7156 – 48.8704	61.8472 – 64.5275	HGtestD3/9
HGsftD3_cot (CoT)	53.4235 – 57.4541	71.1226 – 73.6063	HGtestD3/9
HGsftD3 (Non-CoT)	32.3721 – 36.3752	51.3309 – 54.1694	HGtestD0/1/2
HGsftD3_cot (CoT)	41.5264 – 45.5771	62.4038 – 65.0115	HGtestD0/1/2

3.3 Significance Testing

Setup. We evaluate ablations against a *control* model using paired bootstrap (B=1,000). For ablations that are expected to *decrease* performance (Removal of Basic Formulas `woBF`, Removal of Basic Information `woBI`, Removal of Completion Task `woCP`), we test a one-sided hypothesis $H_0 : \Delta \geq 0$ vs. $H_1 : \Delta < 0$, where Δ = experiment − control. For the Chain-of-Thought (`CoT`) enhancement, which is expected to *improve* performance, we test $H_0 : \Delta \leq 0$ vs. $H_1 : \Delta > 0$. Reported p-values come from the bootstrap test, and the last column gives the *one-sided 95% bound* of Δ (LCB for ablations expecting decreases; UCB for enhancements expecting increases).

Table 31: Statistical Significance Tests (Paired Bootstrap, $B = 1{,}000$). p-values correspond to the one-sided tests described above. The last column reports the one-sided 95% bound of Δ = experiment − control (LCB for expected decreases; UCB for expected increases).

Metric	**p-value**	**One-sided 95% bound of Δ (exp−ctrl)**
Ablation: Removal of Basic Formulas (woBF)		
BLEU-4	0.0010	−4.3816
ROUGE-1	0.0010	−6.2240
ROUGE-2	0.0010	−6.4324
ROUGE-L	0.0010	−6.6014
Ablation: Removal of Basic Information (woBI)		
BLEU-4	0.0010	−3.7227
ROUGE-1	0.0010	−5.5139
ROUGE-2	0.0010	−5.7150
ROUGE-L	0.0010	−5.8678
Ablation: Removal of Prescription Completion Task (woCP)		
BLEU-4	0.0010	−1.5400
ROUGE-1	0.0010	−4.9039
ROUGE-2	0.0010	−0.7661
ROUGE-L	0.0010	−6.0974
Enhancement: Addition of Chain-of-Thought (CoT)		
BLEU-4	0.0010	+4.0926
ROUGE-1	0.0010	+7.9458
ROUGE-2	0.0010	+6.8716
ROUGE-L	0.0010	+8.8137

3.4 Error Analysis for Task 3

To gain qualitative insights into the strengths and limitations of our fine-tuned models on Task 3 (personalized prescription generation), we conducted a structured error analysis on the predictions from the HGtestD0/1/2 test set. Following our statistical analysis plan, for each of the three models fine-tuned on HGsftD0, HGsftD1, and HGsftD2 respectively, we selected the top-3 cases with the highest BLEU-4 scores as high-quality examples and the bottom-3 cases with the lowest BLEU-4 scores as

failure cases. Each case is presented in a standardized format below, including a summary of the input, the reference prescription, the model's output, error tags, and a brief comment.

3.4.1 High-Quality Examples

These examples demonstrate the model's ability to accurately perform personalized modifications to the basic formula based on the patient's specific symptoms, diagnosis, and constitution, resulting in prescriptions highly aligned with the physician's reference.

Example 1 (SFT Data: `HGsftD2`, BLEU-4: 0.831)

Input (summary): 46-year-old female with asthma. Symptoms include nighttime cough and wheezing, history of Hashimoto's thyroiditis. Pulse is wiry and tight at the chi position, weak on deep palpation. Basic formula: Guizhi Qushaoyao Jia Mahuang Xixin Fuzi Tang.

Reference (set): 麻黄、蝉蜕、桂枝、赤芍、干姜、细辛、五味子、半夏、生姜、附子、紫菀、款冬花、甘草、大枣、海螵蛸、当归、枸杞、淫羊藿、补骨脂、菟丝子。

Model output (set): 麻黄、蝉蜕、桂枝、赤芍、干姜、细辛、五味子、半夏、生姜、附子、紫菀、款冬花、补骨脂、核桃仁、白果、甘草、大枣、巴戟天、淫羊藿、菟丝子。

Error tags: None significant.

Comment: The model perfectly replicates the core herb pairs for asthma and accurately adds herbs to tonify kidney yang (e.g., Psoraleae Fructus, Epimedii Herba) for the underlying deficiency (weak pulse). The substitution of Sepiae Endoconcha, Angelicae Sinensis Radix, Lycii Fructus with Juglandis Semen, Ginkgo Semen is a reasonable and clinically valid alternative for consolidating the lungs and tonifying the kidneys.

Example 2 (SFT Data: `HGsftD1`, BLEU-4: 0.811)

Input (summary): 8-year-old female with post-febrile cough. Presents with frequent cough, phlegm in the throat, and yellow-white nasal discharge. Pulse is wiry and slippery. Basic formula: Ling Gan Wu Wei Jiang Xin Ban Xia Xing Ren Tang.

Reference (set): 半夏、厚朴、茯苓、干姜、五味子、生姜、细辛、旋覆花、白芍、麻黄、苦杏仁、甘草、甜叶菊、紫苏叶、大枣、麦冬。

Model output (set): 半夏、厚朴、茯苓、干姜、五味子、生姜、细辛、旋覆花、白芍、麻黄、苦杏仁、甘草、白芷、蒲公英、桔梗、甜叶菊。

Error tags: Minor redundancy/substitution.

Comment: The model correctly identifies the core formula for phlegm-damp cough. To address the yellow nasal discharge (indicating heat), it substitutes Perillae Folium, Ophiopogonis Radix, Jujubae Fructus with Angelicae Dahuricae Radix, Taraxaci Herba (to clear heat and unblock the orifices) and Platycodi Radix (to direct the action upward and expel phlegm), which is a logical and appropriate modification.

Example 3 (SFT Data: `HGsftD0`, BLEU-4: 0.810)

Input (summary): 40-year-old female with post-COVID cough. Presents with severe wheezing and a dry throat without phlegm. Pulse information was not recorded. Basic formula: San Ao Tang.

Reference (set): 旋覆花、白芍、干姜、细辛、五味子、半夏、生姜、紫菀、款冬花、蝉蜕、麻黄、苦杏仁、甘草。

Model output (set): 旋覆花、白芍、半夏、生姜、麻黄、苦杏仁、蝉蜕、干姜、细辛、五味子、代赭石、百部、紫菀、款冬花、甘草。

Error tags: Minor redundancy.

Comment: The model accurately reproduces all core herbs and adds Haematitum to strongly direct qi downward (for severe wheezing) and Stemonae Radix to moisten the lungs and stop cough (for dry throat). These additions are highly appropriate for the specific symptom complex.

3.4.2 Failure Cases

These cases highlight critical failure modes, including the omission of essential herbs, the introduction of irrelevant or contradictory herbs, and a breakdown in holistic reasoning.

Example 1 (SFT Data: `HGsftD0`, BLEU-4: 0.166)

Input (summary): 45-year-old female with epigastric fullness and post-COVID cough. Menstrual flow is scanty, and her basal body temperature chart shows a slow rise in the high-temperature phase. Basic formulas: Sini Tang, Xiongqiong Tang.

Reference (set): 包含大量温阳（附子、干姜）、补肾（淫羊藿、巴戟天、菟丝子）、调经（当归、川芎、益母草）及和胃（百合、乌药、海螵蛸）的药物。

Model output (set): 旋覆花、代赭石、半夏、生姜、党参、甘草、大枣、厚朴、紫苏子、茯苓、款冬花、紫菀、木蝴蝶、干姜、细辛、五味子。

Error tags: Severe omission, Logical inconsistency.

Comment: The model completely ignores the key gynecological issues (scanty menses, slow BBT rise) which point to a core pattern of kidney yang deficiency. Consequently, it omits all yang-tonifying and menstruation-regulating herbs. The output is a generic formula for cough and epigastric fullness, failing to address the patient's holistic condition and demonstrating a critical breakdown in TCM's holistic diagnostic principle.

Example 2 (SFT Data: `HGsftD1`, BLEU-4: 0.241)

Input (summary): 30-year-old female with blood strangury (urinary tract infection, cervicitis, post-miscarriage). Presents with yellow vaginal discharge and a *Gardnerella* infection. Basic formula: Yiyi Fuzi Baijiang San.

Reference (set): 包含清热利湿（败酱草、土茯苓、黄柏）、健脾益气（党参、黄芪）、活血化瘀（当归、赤芍）及外用杀虫止痒药。

Model output (set): 醋延胡索、盐黄柏、薏苡仁、醋五灵脂、白茅根、芡实、茯苓、瞿麦、附子、桑螵蛸、甘草、砂仁、虎杖、牡丹皮、泽泻、山药、熟地黄。

Error tags: Severe omission, Logical inconsistency.

Comment: The model correctly includes some heat-clearing and dampness-draining herbs but commits a critical error by completely omitting **Patriniae Herba**, the key herb from the basic formula. Furthermore, it adds numerous yin- and blood-tonifying herbs (e.g., Rehmanniae Radix Praeparata, Sepiae Endoconcha), which is contraindicated in a condition of damp-heat. This "trapping the pathogen" error shows a fundamental misunderstanding of the disease mechanism.

Example 3 (SFT Data: `HGsftD2`, BLEU-4: 0.057)

Input (summary): 40-year-old female with post-COVID cough. Presents with severe wheezing and a dry throat without phlegm. Pulse information was not recorded. Basic formula: San Ao Tang.

Reference (set): 旋覆花、白芍、干姜、细辛、五味子、半夏、生姜、紫菀、款冬花、蝉蜕、麻黄、苦杏仁、甘草。

Model output (set): （空）

Error tags: Complete omission.

Comment: The model fails to generate any output, resulting in an empty prescription. This catastrophic failure may be due to the missing pulse information in the input, highlighting a potential lack of robustness in handling incomplete clinical data.

4 Methods

4.1 Incremental pre-training

Our study employs autoregressive language modeling (ARLM) as the pre-training task, aiming to predict the next word by maximizing the joint likelihood of the conditional probability sequence. Specifically, for each position t in the input sequence, the model predicts the probability distribution of the current word x_t based on the preceding words $x_1, x_2, ..., x_{t-1}$. The loss function is defined as the mean of the negative log-likelihood:

$$\mathcal{L}_{\text{ARLM}} = -\frac{1}{|D|}\sum_{i=1}^{|D|}\sum_{t=1}^{T_i}\log p_\theta(x_{i,t}|x_{i,1}, ..., x_{i,t-1}) \tag{1}$$

Among them, D represents the training dataset, $|D|$ denotes the total number of samples in the dataset, $x_i = (x_{i,1}, .., x_{i,T_i})$ is the word sequence of the i-th sample (with a length of T_i), and p_θ represents the conditional probability distribution output by the language model with parameters θ. The term $\log p_\theta(x_{i,t}|x_{i,1}, ..., x_{i,t-1})$ denotes the logarithmic probability of the correct prediction for the i-th word in the t-th sample.The loss function minimizes the negative log-likelihood across all samples, forcing the model to learn the dependencies between words within a sequence, thereby enhancing its ability to model TCM language patterns. In the implementation, the model maps the hidden state vector $h_{i,t} \in \mathbb{R}^d$ to the vocabulary space through the classification layer parameters $W \in \mathbb{R}^{d\times K}, b \in \mathbb{R}^K$ (where K is the vocabulary size). The probability distribution is then generated via the Softmax function:

$$p_\theta(x_{i,t} = k|x_{i,<t}) = \frac{\exp(h_{i,t}^T W_k + b_k)}{\sum_{j=1}^{K}\exp(h_{i,t}^T W_j + b_j)} \tag{2}$$

where W_k and b_k represent the weight vector and bias corresponding to the k-th word in the vocabulary. By optimizing the aforementioned loss function, the model iteratively updates the parameters θ, ultimately capturing fine-grained linguistic patterns in the TCM text, thereby laying the foundation for subsequent TCM-specific fine-tuning.

This study conducted incremental pre-training on a distributed GPU cluster using eight NVIDIA A800 GPUs as the hardware configuration. To enhance training efficiency and effectively manage memory, the 8bit AdamW optimizer was employed, which excels in memory management and accelerating training, making it particularly suitable for large-scale model optimization. Additionally, to further reduce memory consumption, BF16 precision training was utilized.

The initial learning rate was set to 3×10^{-5}, adjusted based on the maximum learning rate in the model' s pre-training phase, and dynamically scaled according to batch size, following the principle that the learning rate is proportional to the square

root of the batch size multiplier. To prevent gradient explosion or instability in the early training stage, a warmup strategy was applied, gradually increasing the learning rate to ensure stable training. Specifically, the warmup ratio was set between 0.05 and 0.3, with larger learning rates paired with higher warmup ratios effectively mitigating early-stage instability.

This study also adopted the cosine learning rate decay strategy, which is well-suited for incremental pre-training, as it smoothly decreases the learning rate to prevent premature convergence or gradient oscillations. To further enhance training efficiency, FlashAttention-2 was implemented—an efficient attention computation algorithm that significantly reduces GPU memory usage and accelerates computation, especially beneficial for handling long-sequence inputs in large-scale models. The integration of FlashAttention-2 in the incremental pre-training process effectively improved computational efficiency and optimized the attention mechanism in Transformer models.

The training duration varied depending on the model' s size and architecture. All models were trained under BF16 precision to minimize memory consumption. Furthermore, gradient accumulation was employed to increase throughput, and during the debugging phase, the learning rate and warmup ratio combinations were dynamically adjusted to ensure stability and efficiency throughout the incremental pre-training process.

4.2 Supervised fine-tuning

Specifically, the optimization objective of SFT is to minimize the task-related supervised loss function, which is defined as:

$$\mathcal{L}_{\mathrm{SFT}} = -\frac{1}{|D|} \sum_{(x,y)\in D} \sum_{t=1}^{T} \log p_\theta \left(y_t \mid x, y_{<t}\right) \tag{3}$$

where D represents the domain-annotated dataset, $|D|$ denotes the total number of samples, each consisting of an input x (such as a medical query or instruction) and a target output $y = (y_1, y_2, ..., y_T)$ (such as a standard medical response). $p_\theta \left(y_t \mid x, y_{<t}\right)$ represents the conditional probability of predicting the t-th word based on the input x and the previously generated word sequence $y_{<t}$, given the model parameters θ. By minimizing the negative log-likelihood loss, the model progressively aligns with the data distribution of the TCM task while retaining the language generation capabilities acquired during pre-training.

To improve training efficiency and reduce computational and storage overhead, we employed a parameter-efficient fine-tuning strategy that integrates Low-Rank Adaptation (LoRA) with the PISSA initialization technique. In our approach, LoRA reduces

the scale of newly introduced parameters by decomposing model weights into a low-rank structure, with all layers being adapted to ensure comprehensive modifications across the entire weight space. Specifically, the rank for low-rank decomposition was set to 16, and a scaling factor (Alpha) of 32 was used to balance update magnitude and training efficiency, while a dropout rate of 0.05 further improved generalization. Additionally, the PISSA initialization strategy was adopted to ensure a well-distributed initialization of the new parameters, thereby enhancing optimization efficiency during fine-tuning. To further optimize the learning process, a grid search was conducted to determine the optimal learning rate range (1e-5 to 5e-6), and the warmup ratio was set between 0.05 and 0.3.

4.3 Evaluation methods

BLEU (Bilingual Evaluation Understudy) is a widely used automatic evaluation metric in natural language processing, primarily measuring the similarity between machine-generated text and human reference text. Its core idea is to quantify text quality by statistically analyzing the overlap of n-grams (sequences of n consecutive words) between the candidate and reference texts. BLEU-4 specifically considers multi-granularity matching from 1-gram (unigrams) to 4-gram (four-word sequences), balancing local lexical accuracy and phrase/syntactic coherence. For example, 4-gram matching effectively reflects whether longer semantic units are correctly generated, while 1-gram ensures sufficient coverage of basic vocabulary.

The calculation of BLEU-4 consists of two key components: modified n-gram precision and the brevity penalty (BP). For each n-gram (with n ranging from 1 to 4), we first count the occurrences of all n-grams in the candidate text, but limit the match count of each n-gram to the maximum occurrence of that n-gram in the reference text (known as truncated count). Then, we calculate the ratio of this sum relative to the total number of n-grams in the candidate text, which is referred to as the modified precision:

$$p_n = \frac{\sum \text{Count}_{\text{clip}}(n\text{-gram})}{\sum \text{Count}_{\text{candidate}}(n\text{-gram})} \tag{4}$$

Here, the numerator $\sum \text{Count}_{\text{clip}}(n\text{-gram})$ represents the match count (after truncation) of all n-grams in the candidate translation, while the denominator

$$\sum \text{Count}_{\text{candidate}}(ns\text{-gram}) \tag{5}$$

is the total number of all n-grams in the candidate translation. After that, we take the geometric mean of the precision for 1-gram to 4-gram and weight it (weights are typically evenly distributed, i.e., $w_n = \frac{1}{4}$), and finally multiply by a penalty factor:

$$\text{BLEU-4} = \text{BP} \cdot \exp\left(\frac{1}{4}\sum_{n=1}^{4} \log p_n\right) \tag{6}$$

Among them, BP (Brevity Penalty) is used to punish short candidate text, which is defined as:

$$\mathrm{BP} = \begin{cases} e^{(1-r/c)} & \text{if } c \leq r, \\ 1 & \text{otherwise.} \end{cases} \tag{7}$$

Here, c represents the length (in words) of the candidate text, and r is the length of the reference text.

When the candidate text is significantly shorter than the reference, the BP value reduces the overall score, preventing artificially high precision through overly concise outputs. By incorporating multi-granularity n-gram evaluation, BLEU-4 balances lexical coverage with structural coherence and remains one of the most fundamental and widely used automatic evaluation tools for text generation tasks today.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a collection of metrics used to evaluate text summarization and machine translation quality. Its core principle is to measure the coverage completeness of key information by assessing the overlap between generated and reference texts, focusing on recall-oriented evaluation. Unlike BLEU, which emphasizes precision, ROUGE focuses on how much of the reference content is effectively captured by the generated text, making it particularly suitable for tasks prioritizing information retention (e.g., summarization). The most commonly used variants, ROUGE-1, ROUGE-2, and ROUGE-L, evaluate text from different levels: lexical, phrase, and semantic structure.

ROUGE-1 measures the basic recall by calculating the overlap ratio of words between the generated text and the reference text. Its formula is:

$$R_{\text{ROUGE-1}} = \frac{\sum \text{Count}_{\text{match}}(\text{unigram})}{\sum \text{Count}_{\text{reference}}(\text{unigram})} \tag{8}$$

The numerator $\sum \text{Count}_{\text{match}}(\text{unigram})$ is the total number of words matched between the generated text and the reference text, while the denominator $\sum \text{Count}_{\text{reference}}(\text{unigram})$ is the total number of all words in the reference text. ROUGE-2 further examines the matching of two-word combinations (bigrams), emphasizing the co-occurrence patterns of consecutive words. Its calculation method is similar, with the only difference being that words are replaced by bigrams:

$$R_{\text{ROUGE-2}} = \frac{\sum \text{Count}_{\text{match}}(\text{bigram})}{\sum \text{Count}_{\text{reference}}(\text{bigram})} \tag{9}$$

These metrics, while capable of capturing local matches, struggle to evaluate semantic coherence. To address this, ROUGE-L incorporates a matching mechanism based on the Longest Common Subsequence (LCS): an LCS is defined as the longest sequence of words, which can be continuous or discontinuous, shared by the generated and reference texts without altering the original order. This metric reflects the capability to retain semantic units by calculating the F-value (a balance between precision

and recall) of the LCS:

$$F_{\text{ROUGE-L}} = \frac{(1+\beta^2) \cdot R_{\text{LCS}} \cdot P_{\text{LCS}}}{R_{\text{LCS}} + \beta^2 \cdot P_{\text{LCS}}} \tag{10}$$

Here, $R_{\text{LCS}} = \frac{\text{LCS length}}{\text{Reference text length}}$ represents the coverage ratio (recall) of the LCS in the reference text, $P_{\text{LCS}} = \frac{\text{LCS length}}{\text{Generated text length}}$ represents the coverage ratio (precision) of the LCS in the generated text, and β is a parameter that adjusts the weight between recall and precision (commonly set to 1, indicating equal importance for both).

ROUGE metrics balance lexical coverage and semantic continuity through multi-level matching mechanisms. Currently, they remain among the most mainstream evaluation benchmarks in natural language processing tasks and are often used alongside BLEU to comprehensively assess text generation quality.

5 The complete diagnosis and treatment process in Traditional Chinese Medicine (TCM)

Traditional Chinese Medicine (TCM) employs a systematic and holistic approach to diagnosis and treatment, deeply rooted in its theoretical framework. The diagnostic process is divided into several key steps, including inspection (望, wàng), listening and smelling (闻, wén), inquiry (问, wèn), and palpation (切, qiè). These steps collectively guide the practitioner in identifying the patient' s pattern of disharmony and formulating an individualized treatment plan. 传统中医采用系统化、整体化的诊疗方法，深植于其理论体系之中。诊疗过程主要分为以下几个关键步骤：望（wàng）、闻（wén）、问（wèn）和切（qiè）。这些步骤共同帮助医生识别患者的病证特征，并制定个性化的治疗方案。

5.1 Four examinations (四诊, sì zhěn)

- Inspection (望, wàng):Practitioners observe the patient' s physical appearance, including complexion, tongue characteristics (shape, color, coating), and body posture. Tongue diagnosis is particularly significant, as it reflects the state of internal organs and the progression of disease. 望诊：医生通过观察患者的外部表现，包括面色、舌象（形态、颜色、苔质）和体态等。其中舌诊尤为重要，因为舌象能够反映内脏功能状态及疾病的发展程度。
- Listening and Smelling (闻, wén):This step involves assessing the patient' s voice, speech patterns, and breathing sounds, as well as detecting any unusual odors from the body, breath, or excretions. 闻诊：医生通过听患者的声音、语调和呼吸音，以及嗅察身体、口腔或排泄物中是否存在异常气味。

- Inquiry (问, wèn):A detailed interview is conducted to gather information about the patient’s symptoms, lifestyle, emotional state, diet, sleep patterns, past medical history, and family history. Special attention is given to the “Ten Questions” (十问, shí wèn), which cover areas such as chills and fever, sweating, appetite, urination, defecation, and menstrual health. 问诊：医生通过详细的问诊收集患者的症状、生活习惯、情绪状态、饮食睡眠情况、既往病史及家族病史等信息。问诊重点包括“十问”，如寒热、汗出、饮食、二便及月经等方面。
- Palpation (切, qiè):Palpation focuses on pulse diagnosis, where the practitioner assesses the pulse at three positions on each wrist (寸, cùn; 关, guān; 尺, chǐ) and at three levels of depth (superficial, intermediate, and deep). The pulse’s rhythm, speed, strength, and quality provide insights into the state of the internal organs and qi (vital energy). 切诊：医生通过把脉观察脉象变化，在患者双腕的三部（寸、关、尺）及三候（浮、沉、中）中感知脉搏的节律、速度、力度及特性，从而判断内脏功能及气血状况。

5.2 Diagnosis and syndrome differentiation (辨证, biàn zhèng)

After collecting data from the four diagnostic methods, the practitioner synthesizes this information to identify a specific pattern of disharmony (证, zhèng). Pattern differentiation is the cornerstone of TCM and involves classifying the patient’s condition based on core theoretical systems, such as: 通过四诊收集信息后，医生综合分析以确定具体的证型（zhèng）。辨证是中医的核心，通过以下理论分类患者的病证状态：

- The Eight Principles (八纲辨证, bā gāng biàn zhèng): Yin-Yang, Interior-Exterior, Cold-Heat, and Deficiency-Excess. 八纲辨证：阴阳、表里、寒热、虚实。
- Zang-Fu Organ Theory (脏腑辨证, zàng fǔ biàn zhèng): Patterns related to the dysfunction of internal organs. 脏腑辨证：与内脏功能失调相关的病证。
- Qi, Blood, and Body Fluids (气血津液辨证, qì xuè jīn yè biàn zhèng): Disorders involving qi stagnation, blood stasis, or fluid imbalance. 气血津液辨证：涉及气滞、血瘀或津液失调的病证。
- Six Stages and Four Levels (六经辨证与卫气营血辨证): Diagnostic systems used for specific diseases such as febrile illnesses. 六经辨证与卫气营血辨证：常用于治疗外感热病的诊断系统。

5.3 Formulating the treatment strategy (治法, zhì fǎ)

Once the pattern is determined, the practitioner formulates a treatment principle aimed at restoring balance. This principle guides the selection of a basic formula (基

础方, jī chǔ fāng), which serves as the treatment’ s foundation. 在确立证型后，医生制定以恢复平衡为目标的治疗原则。该原则指导基础方的选择，为治疗提供核心配方。Example Formulas 实例方剂:

- Si Jun Zi Tang (四君子汤): A formula for spleen qi deficiency. 治疗脾气虚。
- Gui Zhi Tang (桂枝汤): A formula for exterior wind-cold syndrome. 治疗外感风寒表证。

5.4 Common basic formulas and their sources (常用基础方及其来源)

Basic formulas in TCM are derived from classical texts and are designed to treat specific patterns of disharmony. These formulas are the foundation of treatment and are often modified to suit individual needs. Below are some commonly used formulas and their historical origins: 中医中常用的基础方源于经典医籍，旨在针对特定的病证模式进行治疗。这些方剂是治疗的基础，常根据个体需求进行加减调整。以下列举了一些常用基础方及其历史来源：

- Si Jun Zi Tang (四君子汤, Four Gentlemen Decoction)Source: 《太平惠民和剂局方》(Formulary of the Pharmacy Service for Benefiting the People in the Taiping Era). 出处：《太平惠民和剂局方》。Indication: Spleen qi deficiency, characterized by fatigue, poor appetite, and loose stools. 主治：脾气虚证，如疲倦乏力、食欲不振、大便稀溏。
- Gui Zhi Tang (桂枝汤, Cinnamon Twig Decoction)Source: 《伤寒论》(Treatise on Cold Damage), by Zhang Zhongjing. 出处：张仲景《伤寒论》。Indication: Exterior wind-cold syndromes with deficiency, presenting with mild fever, sweating, and aversion to wind. 主治：外感风寒表虚证，症见微热、汗出、恶风。
- Liu Wei Di Huang Wan (六味地黄丸, Six-Ingredient Rehmannia Pill)Source: 《小儿药证直诀》(Key to Therapeutics of Children’s Diseases), by Qian Yi. 出处：钱乙《小儿药证直诀》。Indication: Kidney yin deficiency, characterized by dizziness, tinnitus, and lumbar soreness. 主治：肾阴虚证，症见头晕耳鸣、腰膝酸软。
- Xiao Chai Hu Tang (小柴胡汤, Minor Bupleurum Decoction)Source: 《伤寒论》(Treatise on Cold Damage), by Zhang Zhongjing. 出处：张仲景《伤寒论》。Indication: Shaoyang syndrome, characterized by alternating fever and chills, bitter taste, and hypochondriac pain. 主治：少阳证，症见寒热往来、口苦胁痛。
- Ba Zhen Tang (八珍汤, Eight Treasures Decoction)Source: 《正体类要》(Essentials of Orthodox Medicine). 出处：《正体类要》。oIndication: Qi and blood deficiency, characterized by pale complexion, dizziness, and fatigue. 主治：气血两虚证，症见面色苍白、头晕乏力。

5.5 Implementation, individualization, and follow-up adjustments (实施、个性化治疗与复诊调方)

5.5.1 Individualization of the formula (方剂个性化治疗)

Once the fundamental formula is selected, the practitioner tailors it to the patient's specific needs. This process, known as formula modification (方剂加减, fāng jì jiā jiǎn), involves adding or removing herbs to enhance therapeutic effects or address additional symptoms. 在选定基础方后，医生会根据患者的具体需求对方剂进行个性化调整，即“方剂加减”。这一过程通过增减药物来增强疗效或针对附加症状进行治疗。Example modifications: 加减示例：

- Adding Huang Qi (黄芪) to strengthen qi in cases of severe fatigue. 在严重疲劳时加黄芪以补气。
- Removing Gui Zhi (桂枝) in cases of sweating to avoid excessive warming. 在伴随汗出的情况下去桂枝以避免过度温热。

5.5.2 Adjustments during follow-up visits (复诊调方)

Each follow-up visit provides an opportunity to reassess the patient's progress and adjust the prescription accordingly. The practitioner evaluates changes in symptoms, pulse, tongue characteristics, and overall condition to refine the treatment. 每次复诊为重新评估患者的病情进展及调整方剂提供了机会。医生会通过观察症状变化、脉象、舌象及整体状态，进一步优化治疗方案。Steps in follow-up adjustments:

- Symptom reassessment (症状再评估): Identify improvements or new symptoms that may have arisen. 明确哪些症状已改善，以及是否出现新的症状。
- Pulse and tongue analysis (脉象与舌象分析): Assess changes in pulse strength, rhythm, and tongue coating or color. 脉象与舌象分析：观察脉搏的强弱、节律及舌苔、舌色的变化。
- Adjusting the formula (方剂调整): Modify the dosage or composition of herbs to reflect the patient's current condition. 根据患者的当前状态调整药物剂量或组成。For example: 例如：Reducing heating herbs if symptoms of internal heat appear. 若出现内热症状，减少温热药物。Adding cooling herbs if heat signs persist. 若热象持续，增加清热药物。

Strategic Planning (治疗策略调整): Reevaluate long-term treatment goals and ensure the plan aligns with the patient's recovery trajectory. 治疗策略调整：重新评估长期治疗目标，确保方案与患者的康复进程一致。

6 Terminology

WHO terminology explanation.

- "**Holism**": One of the philosophical ideas regarding the human body as an organic whole, which is integrated with the external environment.
- "**Syndrome differentiation and treatment**": Diagnosis of the syndrome, through comprehensive analysis of symptoms and signs, which has implications for determining the cause, nature and location of the illness and the patient' s physical condition, and their treatment.
- "**Syndrome differentiation**": The process of overall analysis of clinical data to determine the location, cause and nature of a patient' s disease and achieving a diagnosis of a pattern/syndrome, also called pattern/syndrome differentiation.
- "**Different treatments for the same disease**": Applying different methods of treatment to the same kind of disease but have different patterns/syndromes.
- "**Principles, methods, formulas, and medicinals**": The four basic steps of diagnosis and treatment: determining the cause, mechanism and location of the disease according to the medical theories and principles, then deciding the treatment principle and method, and finally selecting a formula as well as proper medicinals.
- "**Four examinations**": A collective term for inspection, listening and smelling, inquiry, and palpation.
- "**Insominia**": Prolonged inability to obtain normal sleep.
- "**Wasting-thrist**": Any diseased state characterized by polydipsia, polyphagia, and polyuria, similar to diabetes.
- "**Running piglet**": An ancient name for the morbid condition characterized by a feeling of masses of gas ascending within the abdomen like running piglets, also known as running piglet qi.
- "**Urticaria**": An allergic disorder of the skin, marked by red or pale wheals, intermittent, associated with intense itching.
- "**Spotting**": Slight but persistent leakage of blood from the uterus, the same as metrostaxis.
- "**Infertility**": Lack of capacity to produce offspring.
- "**Delayed Menstruation**": Periods that come one week or more after due time, for more than two successive periods.
- "**Lesser Yang Disease Pattern**": A pattern/syndrome in which the pathogen exists between the exterior and interior of the body, marked by alternate fever

and chills, fullness and choking feeling in the chest and hypochondriac region, dry throat and string-like pulse, also called the lesser yang disease.

- "**Greater Yin Disease Pattern**":A pattern/syndrome characterized by decline of spleen yang with production of cold-dampness, and manifested by anorexia, vomiting, abdominal fullness and dull pain, diarrhea and weak pulse, also called the greater yin disease.
- "**Mammary hyperplasia**": Benign hyperplasia of mammary gland.

Terminology explanation by ourself.

- "**Cough due to External Contraction**": Cough due to External Contraction (外感咳嗽, Wài Gǎn Ké Sou) refers to a sudden-onset cough caused by invasion of external pathogens (wind, cold, heat, or dryness). It is commonly seen in acute respiratory infections (e.g., common cold, flu, bronchitis) and is differentiated by the nature of the pathogen and the body' s response.
- "**Throat Impediment**": Throat Impediment (喉痹, Hóu Bì) in Traditional Chinese Medicine (TCM) refers to a syndrome characterized by obstruction, swelling, or discomfort in the throat, often accompanied by pain, dryness, difficulty swallowing, or a sensation of blockage.
- "**Stomach duct pain**": Stomach duct pain (胃脘痛, Wèi Wǎn Tòng) refers to recurrent or episodic pain in the epigastric region (the area between the ribs and navel, centered around the stomach). It is a hallmark symptom of digestive dysfunction in TCM, encompassing conditions like gastritis, ulcers, or functional dyspepsia but analyzed through TCM pattern differentiation.
- "**Sinusitis**": Sinusitis (鼻渊, Bí Yuān) refers to a chronic or recurrent condition characterized by turbid nasal discharge, congestion, headaches, and impaired smell, often linked to sinus infections, allergies, or nasal inflammation in Western medicine. However, TCM interprets it as a dysfunction of the Lung, Spleen, or Gallbladder systems, often caused by wind-heat, damp-heat, or deficiency patterns.
- "**Gastric distention disorder**": Gastric distention disorder (胃痞病, Wèi Pǐ Bìng) refers to a chronic or recurrent sensation of fullness, bloating, and discomfort in the epigastric region (upper abdomen) without actual pain. It is often associated with digestive dysfunction, emotional stress, or spleen-stomach imbalances and differs from "stomach duct pain" (胃脘痛) by the absence of sharp pain.
- "**Fever from External Contraction**": Fever from External Contraction (外感发热病, Wài Gǎn Fā Rè Bìng) refers to acute febrile conditions caused by invasion of external pathogens, primarily characterized by sudden onset of fever

with accompanying exterior symptoms. This condition represents the body's defensive response to pathogenic factors and is commonly seen in viral/bacterial infections, influenza, or early-stage infectious diseases.

- "**Rhinitis**": Rhinitis (鼻噫病, Bí Tì Bìng) in TCM refers to chronic or recurrent nasal discharge, sneezing, and congestion, primarily caused by lung-spleen deficiency, wind-cold invasion, or internal damp-phlegm accumulation. Unlike Western medicine's focus on allergens/infection, TCM emphasizes systemic imbalances that make the nose vulnerable to pathogens.
- "**Acne**": Acne (痤疮, Cuó Chuāng) is a heat-toxin disorder primarily affecting the face, chest, and back, characterized by inflammatory papules, pustules, or nodules due to lung/ stomach heat, blood stasis, or damp-toxin accumulation. Unlike Western dermatology's focus on bacteria/hormones, TCM treats acne as a systemic imbalance manifesting in the skin.
- "**Regulation of Constitution and Sub-health**": TCM Regulation of Constitution and Sub-health (调理, Tiáo Lǐ): A holistic approach using herbal medicine to balance bodily functions, optimize health, and address imbalances causing discomfort or reduced vitality through Yin-Yang adjustment and Qi-Blood circulation.
- "**Chronic throat impediment**": Chronic throat impediment (慢喉痹, Màn Hóu Bì) refers to long-standing throat discomfort (pain, dryness, foreign body sensation) without acute infection, caused by yin deficiency, qi stagnation, or blood stasis. Unlike Acute Throat Impediment (急喉痹), it involves chronic inflammatory patterns often resistant to conventional antibiotics.
- "**Cough due to Internal Damage**": Cough due to Internal Damage (内伤咳嗽, Nèi Shāng Ké Sòu) refers to chronic or recurrent coughing without external pathogen involvement, caused by organ system imbalances (primarily Lung, Spleen, Liver, or Kidney dysfunction). Unlike acute exogenous coughs (外感咳嗽), it is characterized by:Long duration (>8 weeks), Relapsing nature, Association with constitutional weakness.
- "**Trying to Conceive**": Trying to Conceive (备孕, Bèi Yùn) is a holistic approach in TCM that focuses on optimizing fertility by balancing the body's internal environment. It involves regulating the menstrual cycle, enhancing kidney essence (肾精, Shèn Jīng), harmonizing qi and blood (气血, Qì Xuè), and addressing underlying patterns of imbalance such as spleen/stomach weakness, liver qi stagnation (肝气郁结, Gān Qì Yù Jié), or dampness-phlegm accumulation (痰湿, Tán Shī). Unlike Western medicine' s focus on hormonal stimulation or assisted reproductive technologies, TCM emphasizes dietary ad-

justments, herbal therapies, acupuncture, and lifestyle modifications to cultivate reproductive health and improve the chances of conception naturally.

- "**Eczema**": Eczema (湿疹, Shī Zhěn) is a chronic skin condition characterized by itchy, red, and inflamed skin due to heat, dampness, and wind pathogens in the body, according to TCM. It often manifests as papules, vesicles, or exudative lesions primarily on the flexor surfaces of the body. TCM views eczema as a result of spleen and stomach dysfunction leading to damp-heat accumulation, coupled with external wind invasion that exacerbates itching. Unlike Western medicine's approach focusing on suppressing symptoms with corticosteroids, TCM treats eczema by removing dampness, clearing heat, expelling wind, and nourishing blood to alleviate symptoms and address the root cause.
- "**Wenbing with Damp-Heat Pattern**": Wenbing with Damp-Heat Pattern (湿温病, Shī Wēn Bìng) is a TCM syndrome characterized by the invasion of external dampness and heat pathogens, often occurring during humid seasons. It manifests as fever, heavy-headedness, chest oppression, fatigue, nausea, and sticky sweat due to impaired spleen function and stagnation of damp-heat in the middle and lower jiao. Unlike acute febrile diseases treated with antibiotics or antivirals in Western medicine, TCM addresses this condition by promoting the transformation and elimination of dampness, clearing heat, and restoring qi transformation to reestablish balance. Herbal formulas like San Ren Tang are commonly used to target this pattern.
- "**Cough**": Acute Throat Impediment (急喉痹, Jí Hóu Bì) is a TCM condition characterized by sudden throat pain, swelling, and difficulty swallowing due to external wind-heat or wind-cold invasion obstructing the throat. It often presents with symptoms like hoarseness, redness of the throat, and fever. TCM views this as an acute blockage of qi and blood flow caused by pathogenic factors attacking the lung and throat. Treatment focuses on dispelling wind, clearing heat, resolving toxicity, and reducing inflammation, often using herbs like Jin Yin Hua (金银花) and Bo He (薄荷), unlike Western medicine' s reliance on antibiotics or anti-inflammatory drugs.

7 Abbreviations

Abbrev.	Definition
LLM	Large Language Model
TCM	Traditional Chinese Medicine
ipt	incremental pre-training
sft/SFT	supervised fine-tuning
CoT	Chain-of-Thought
GAM	Guang'anmen Hospital (Chinese Academy of Traditional Chinese Medicine)
HG	Han'gu TCM Clinic
BD	inferring the Basic formulas from Diagnosis
BF	Basic Formula
BI	Basic Information
BR	Basic Information from prescription Results
CP	Completing Prescriptions
DP	Diagnosis from Prescriptions
BLEU-4	Bilingual Evaluation Understudy (4-gram)
ROUGE-1/2/L	Recall-Oriented Understudy for Gisting Evaluation (1-gram / 2-gram / LCS-based)

8 Unified Notation for BLEU/ROUGE

Table 33: Unified notation for evaluation metrics (BLEU-4 and ROUGE series)

Symbol	Domain	Definition / Notes
c	$\mathbb{N}$	Length (number of tokens/words) of the *candidate* sequence.
r	$\mathbb{N}$	Length of the *reference* sequence.
BP	$[0,1]$	**Brevity Penalty**: $\mathrm{BP} = \begin{cases} 1, & c > r \\ \exp\left(1 - \frac{r}{c}\right), & c \le r \end{cases}$
p_n	$[0,1]$	Modified n-gram precision (order n; clipped counts).
N	$\{1,2,3,4\}$	Max n for BLEU; we use N=4 (BLEU-4).
$\mathrm{LCS}(X,Y)$	$\mathbb{N}$	Length of the Longest Common Subsequence (candidate X, reference Y).
R_{LCS}	$[0,1]$	ROUGE-L recall: $R_{\mathrm{LCS}} = \mathrm{LCS}/r$.
P_{LCS}	$[0,1]$	ROUGE-L precision: $P_{\mathrm{LCS}} = \mathrm{LCS}/c$.
β	$[0,\infty)$	Balance parameter in ROUGE-L: $F_{\mathrm{ROUGE\text{–}L}} = \dfrac{(1+\beta^2)\, R_{\mathrm{LCS}}\, P_{\mathrm{LCS}}}{R_{\mathrm{LCS}} + \beta^2 P_{\mathrm{LCS}}}$.

We report BLEU-4 as $\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\frac{1}{4}\sum_{n=1}^{4} \log p_n\right)$. ROUGE-1/2 are recalls; ROUGE-L uses F-score with default β=1 (F1).

Table 34: Evaluation metric configuration used in this study

Item	**Setting**
BLEU	BLEU-4 with N=4; BLEU = BP $\cdot$ $\exp\left(\frac{1}{4}\sum_{n=1}^{4}\log p_n\right)$; BP as in Table 33.
ROUGE-1/2	Unigram and bigram *recall*, respectively.
ROUGE-L	F-score form with β=1 (F1).
Reporting	Per-task scores computed on test sets, then **averaged across the six tasks**.

9 Related Works

Study/Model	Model & Training	Tasks	Data Source	Key Contribution
This work	Qwen2.5-7B-Instruct; two-stage training; 10% CoT; physician-task-specific SFT; ablations; compared against 6 LLMs	Diagnosis; basic formula construction; personalized prescription generation	100,538 clinical records	Basic formula matters; 10% CoT $\approx$ +20%; generalizes to unseen school; quantifies cross-physician basic formula variation (threshold 0.5).
TCMKG ([1])	DL TCM knowledge graph; NER + relation extraction to build KG	KG retrieval; visualization	TCM literature; classic formula corpora	Early DL-powered TCM-KG enabling structured query and visual exploration
Meridian Prediction ([2])	Classical ML (RF/SVM/DT/kNN); molecular fingerprints + ADME; cross-validation	Meridian classification	TCMID (646 herbs; 10,053 compounds)	Links chemical features to meridian labels; RF balanced accuracy up to 0.83
Lingdan ([3])	Baichuan2-13B LLM; continued pre-train + QLoRA SFT; multi-task	TCM QA; clinical reasoning; prescription recommendation	Classics; textbooks; EMRs; pharmacopoeia	"Knowledge-to-language" encoding; multi-turn diagnostic dialogues (TCM-IDDF)
Zhongjing ([4])	Ziya-LLaMA LLM; continued pre-train + SFT + PPO-RLHF	Medical QA; multi-turn consultation	Textbooks; real-world dialogues; CMtMedQA	Expert feedback + reward models for safety/alignment
ChiMed-GPT ([5])	Ziya-13B-v2; CMD pre-train + ChiMed SFT + RLHF; 4k context	Medical QA; dialogue	TCM encyclopedias; ChiMed; MedDialog	Full training regime; long-context consultation
Qibo ([6])	Chinese-LLaMA 7B/13B; 2GB TCM corpus for continued pre-train + full FT	QA; retrieval-augmented consultation	Classics; prescriptions; ChatMed-TCM; CMtMedQA	Qibo-Benchmark; retrieval-based consultation + syndrome differentiation
MedChatZH ([7])	Baichuan-7B; TCM classics pre-train + full instruction tuning	Consultation QA (TCM + WM)	TCM classics; curated instruction sets	Bilingual consultation; safety-aware instruction tuning
HPA-UNet ([8])	Improved U-Net; hybrid post-processing attention + augmentation	Tongue image segmentation	BioHit tongue images	Post-processing attention improves boundary detail and robustness
Chaos-MLP ([9])	Chaotic-transform MLP-like; chaotic feature fusion + bi-centroid loss	Multi-label body constitution recognition	Facial/tongue datasets (MFBC; MTBC)	Chaos-based MLP pipeline; dual datasets
SDPR ([10])	Multi-task graph; four-partite graph + syndrome-induced pre-train + therapy-aware contrastive learning	Prescription recommendation (with syndrome differentiation)	Public + real prescriptions	Explicit symptom $\rightarrow$ syndrome $\rightarrow$ therapy $\rightarrow$ herb chain
NFFGRAM ([11])	Deep fusion network; nonlinear multi-feature fusion + gated recurrent self-attention	Formula recommendation	Formula + patient features	Models complex patient–formula relations under sparsity

Table 35: Representative works at the intersection of TCM and AI.

References

[1] Z. Zheng, Y. Liu, Y. Zhang, and C. Wen, "TCMKG: A deep learning based traditional Chinese medicine knowledge graph platform," in *ICKG*, pp. 560–564, 2020.

[2] Y. Wang, M. Jafari, Y. Tang, and J. Tang, "Predicting Meridian in Chinese traditional medicine using machine learning approaches," *PLoS Comput. Biol.*, vol. 15, p. e1007249, 2019.

[3] R. Hua, X. Dong, Y. Wei, Z. Shu, P. Yang, Y. Hu, S. Zhou, H. Sun, K. Yan, X. Yan, *et al.*, "Lingdan: enhancing encoding of traditional Chinese medicine knowledge for clinical reasoning tasks with large language models," *J. Am. Med. Inform. Assn.*, vol. 31, pp. 2019–2029, 2024.

[4] S. Yang, H. Zhao, S. Zhu, G. Zhou, H. Xu, Y. Jia, and H. Zan, "Zhongjing: Enhancing the Chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue," in *AAAI*, pp. 19368–19376, 2024.

[5] Y. Tian, R. Gan, Y. Song, J. Zhang, and Y. Zhang, "ChiMed-GPT: A Chinese Medical Large Language Model with Full Training Regime and Better Alignment to Human Preferences," in *ACL*, pp. 7156–7173, 2024.

[6] H. Zhang, X. Wang, Z. Meng, Z. Chen, P. Zhuang, Y. Jia, D. Xu, and W. Guo, "Qibo: A large language model for traditional Chinese medicine," *arXiv preprint arXiv:2403.16056*, 2024.

[7] Y. Tan, Z. Zhang, M. Li, F. Pan, H. Duan, Z. Huang, H. Deng, Z. Yu, C. Yang, G. Shen, *et al.*, "MedChatZH: A tuning LLM for traditional Chinese medicine consultations," *Comput. Biol. Med.*, vol. 172, p. 108290, 2024.

[8] L. Yao, Y. Xu, S. Zhang, J. Xiong, A. Shankar, M. H. Abidi, and M. Nappi, "HPA-UNet: A Hybrid Post-Processing Attention U-Net for Tongue Segmentation," *IEEE J. Biomed. Health.*, 2024.

[9] M. Zhang, G. Wen, P. Yang, C. Wang, X. Huang, and C. Chen, "Chaos-MLP: Chaotic Transform MLP-like Architecture for Medical Images Multi-label Recognition Task," *IEEE J. Biomed. Health.*, 2024.

[10] W. Yue, W. Ji, X. Wang, X. Ma, P. Wang, and X. Wang, "SDPR: Prescription Recommendation with Syndrome Differentiation in Traditional Chinese Medicine," *IEEE J. Biomed. Health.*, 2025.

[11] H. Hu, Y. Li, and Z. Li, "NFFGRAM: Nonlinear Multi-feature Fusion and Gated Recurrent Self-Attention Mechanism for Traditional Chinese Medicine Formula Recommendation," *IEEE J. Biomed. Health.*, 2025.

10 Appendix

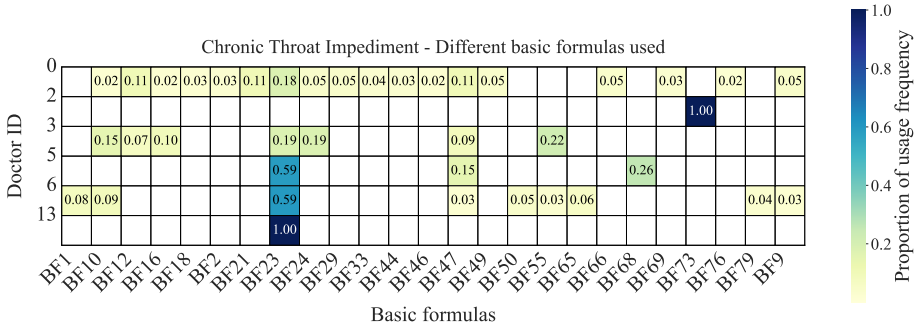

Fig. 3: Variation heatmap of Chronic Throat Impediment.

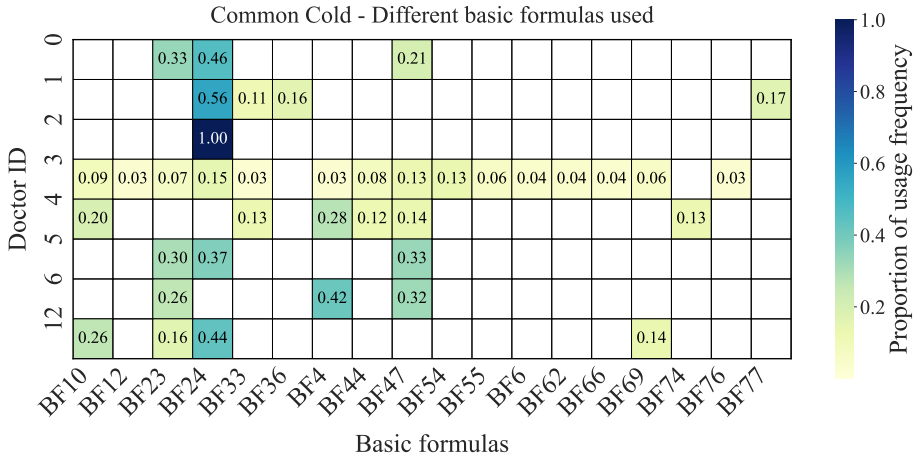

Fig. 4: Variation heatmap of Common Cold.

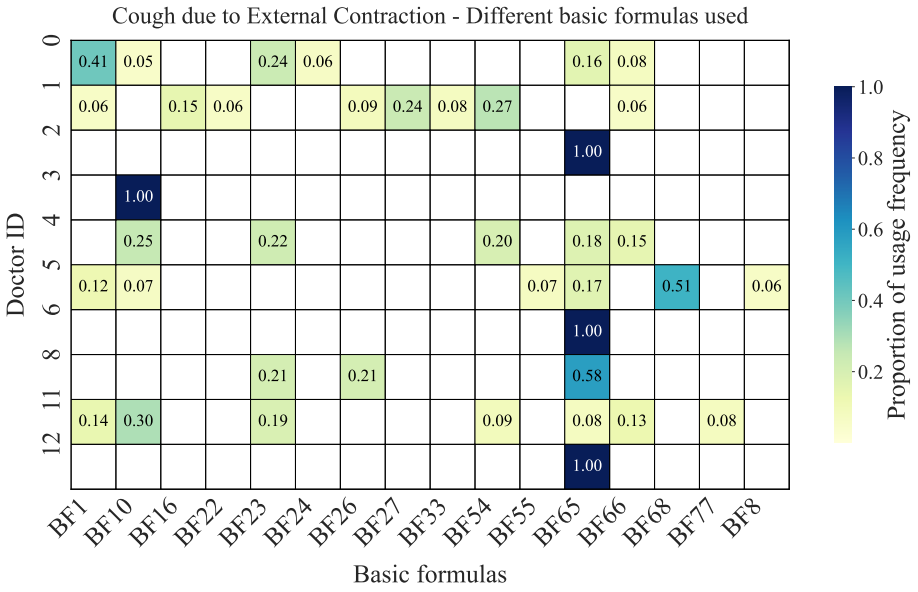

Fig. 5: Variation heatmap of Cough due to External Contraction.

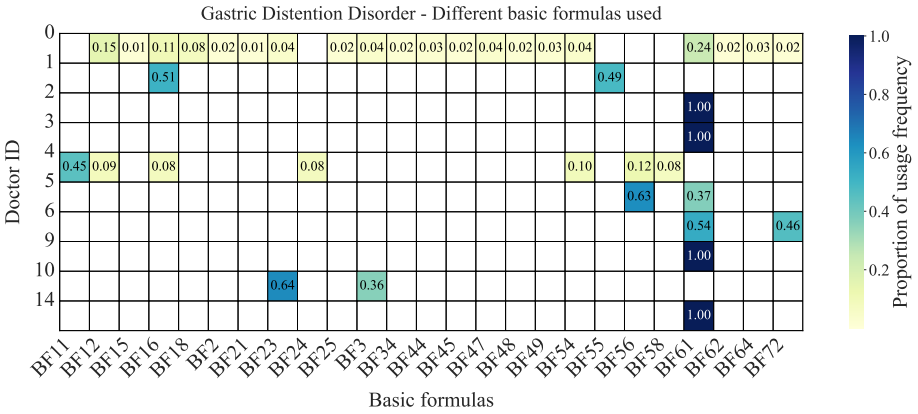

Fig. 6: Variation heatmap of Gastric Distention Disorder.

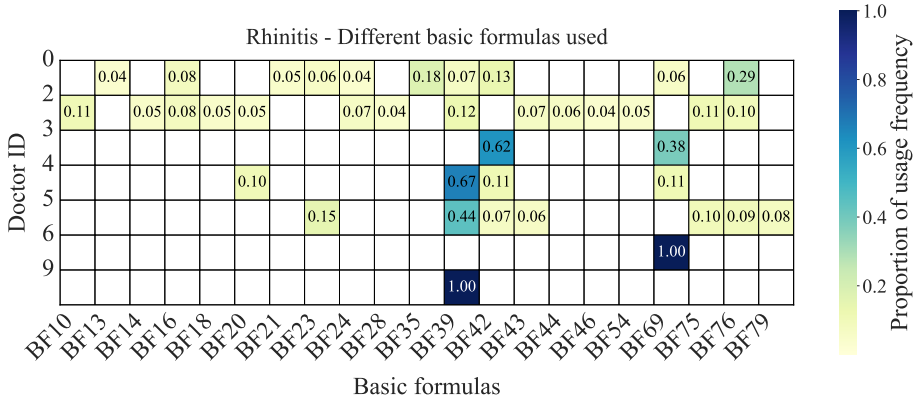

Fig. 7: Variation heatmap of Rhinitis.

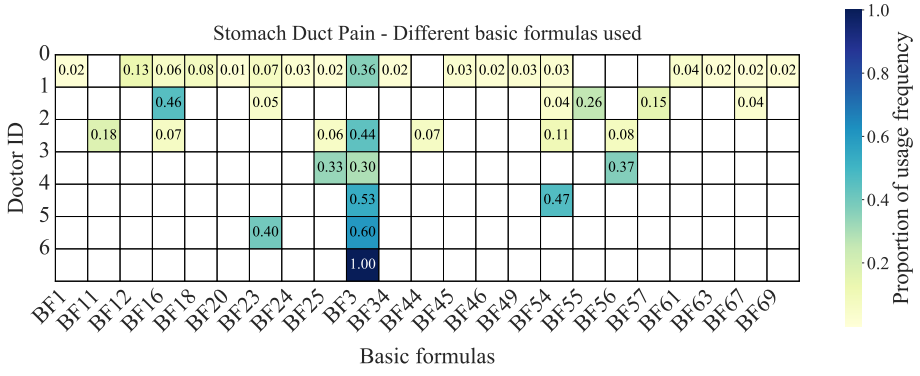

Fig. 8: Variation heatmap of Stomach Duct Pain.

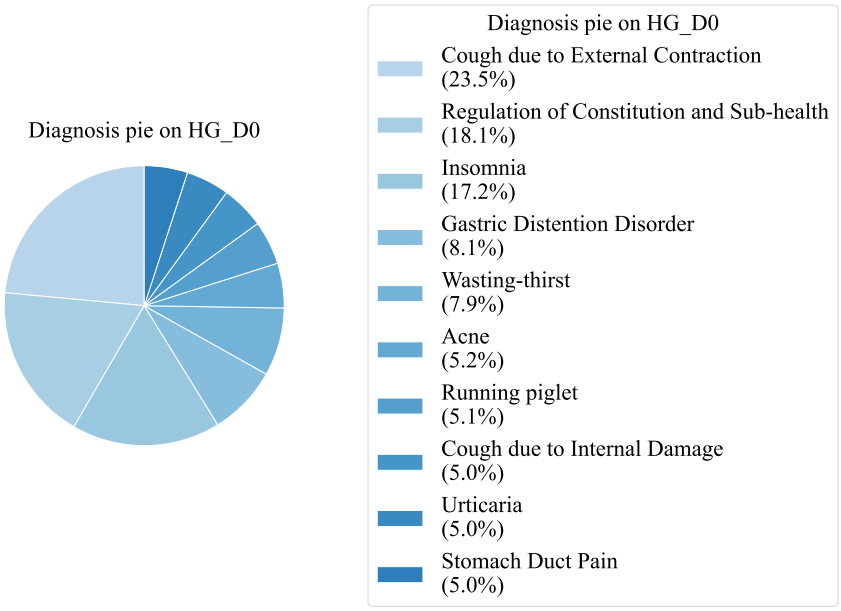

Fig. 9: Diagnosis pie on HG_D0

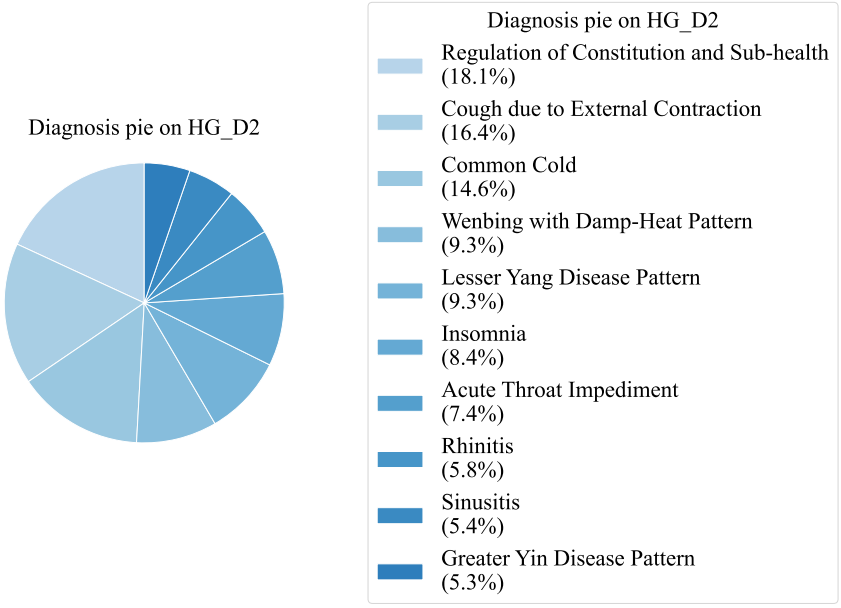

Fig. 10: Diagnosis pie on HG_D2